airbnb in New York City

Carole Mattmann und Jonas Zuercher

13 Februar 2020

Included packages:

```r
library(dplyr)
library(tidyverse)
library(geosphere)
library(ggplot2)
```

Introduction

We are exploring a dataset of airbnb listings in New York City in 2019.

```r
AB_NYC <- read.csv("../01_data/AB_NYC_2019.csv", header=TRUE)
str(AB_NYC)
```

```
## 'data.frame':    48895 obs. of  16 variables:
##  $ id                            : int  2539 2595 3647 3831 5022 5099 5121 5178 5203 5238 ...
##  $ name                          : Factor w/ 47906 levels "","'Fan'tastic",..: 12661 38172 45171 1570
##  $ host_id                       : int  2787 2845 4632 4869 7192 7322 7356 8967 7490 7549 ...
##  $ host_name                     : Factor w/ 11453 levels "","'Cil","-TheQueensCornerLot",..: 5051 48
##  $ neighbourhood_group           : Factor w/ 5 levels "Bronx","Brooklyn",..: 2 3 3 2 3 3 2 3 3 3 ...
##  $ neighbourhood                 : Factor w/ 221 levels "Allerton","Arden Heights",..: 109 128 95 42
##  $ latitude                      : num  40.6 40.8 40.8 40.7 40.8 ...
##  $ longitude                     : num  -74 -74 -73.9 -74 -73.9 ...
##  $ room_type                     : Factor w/ 3 levels "Entire home/apt",..: 2 1 2 1 1 1 2 2 2 1 ...
##  $ price                         : int  149 225 150 89 80 200 60 79 79 150 ...
##  $ minimum_nights                : int  1 1 3 1 10 3 45 2 2 1 ...
##  $ number_of_reviews             : int  9 45 0 270 9 74 49 430 118 160 ...
##  $ last_review                   : Factor w/ 1765 levels "","2011-03-28",..: 1503 1717 1 1762 1534 17
##  $ reviews_per_month             : num  0.21 0.38 NA 4.64 0.1 0.59 0.4 3.47 0.99 1.33 ...
##  $ calculated_host_listings_count: int  6 2 1 1 1 1 1 1 1 4 ...
##  $ availability_365              : int  365 355 365 194 0 129 0 220 0 188 ...
```

Data cleaning

Data import and preparation of the Airbnb Dataset

Following changes have been made to the dataset:

remove price 0

remove all listings with price 0

```r
AB_NYC <-AB_NYC[AB_NYC$price > 0,]
```

add log price

add logarithmic price for analysis purposes

```
AB_NYC <- cbind(AB_NYC,price_log = log(AB_NYC$price))
```

remove inactive listings

remove inactive listings and make new dataset to compare to full dataset

```
AB_NYC_available <- AB_NYC %>%
  filter(availability_365 > 0)
```

add distance to Times Square to model

We want to make a statement about how central the place is. Therefore the distance to Times Square is caculated using the latitude and longitude of the listings. The package "geosphere" is used.

Times Square, Manhattan, NY, USA, Latitude and longitude coordinates are: 40.758896, -73.98513

```
coord <- cbind(AB_NYC_available$longitude,AB_NYC_available$latitude)
dist.timessquare <- distGeo(p1=coord, p2=c(-73.985130, 40.758896))
AB_NYC_available <- cbind(AB_NYC_available,dist.timessquare)
```

Create subsets for room type

```

AB_NYC_entirehome <-AB_NYC_available[AB_NYC_available$room_type == "Entire home/apt",]

AB_NYC_privateroom <-AB_NYC_available[AB_NYC_available$room_type == "Private room",]

AB_NYC_sharedroom <-AB_NYC_available[AB_NYC_available$room_type == "Shared room",]
```

Prepare dataset for merging

```
# Neighbourhood Group klein schreiben für Merging
AB_NYC_available$neighbourhood_group<-tolower(AB_NYC_available$neighbourhood_group)

#Leerzeichen aus den Distrikten entfernen
AB_NYC_available$neighbourhood_group <-gsub(" ","", AB_NYC_available$neighbourhood_group)

# Neighbourhood Group als Faktor
AB_NYC_available$neighbourhood_group<-factor(AB_NYC_available$neighbourhood_group)
```

Data import and preparation of the Airbnb Dataset

Das Dataset wurde auf folgender URL heruntergeladen: https://data.cityofnewyork.us/City-Government/Agency-Performance-Mapping-Indicators-Annual/gsj6-6rwm

```
Ind_NYC<- read.csv("../01_data/Indicators_NYC.csv")
head(Ind_NYC)
```

```
##   Agency    Geographic.Unit Geographic.Identifier
## 1    DCA Community District       Staten Island 3
## 2    DCA Community District       Staten Island 2
## 3    DCA Community District       Staten Island 1
## 4    DCA Community District             Queens 14
## 5    DCA Community District             Queens 13
## 6    DCA Community District             Queens 12
##                        Indicator FY2011 FY2012 FY2013 FY2014 FY2015 FY2016
## 1 Resolved Consumer Complaints     44     40     53     38     38     33
## 2 Resolved Consumer Complaints     46     57     56     43     29     63
## 3 Resolved Consumer Complaints     75     56     29     61     42     65
## 4 Resolved Consumer Complaints     17     25      9      8      8     11
## 5 Resolved Consumer Complaints     64     36     22     41     44     61
## 6 Resolved Consumer Complaints    125    144    113    113    112    122
##   FY2017 FY2018 FY2019
## 1     22     29     14
## 2     23     25     26
## 3     46     28     34
## 4     14     23     25
## 5     36     45     40
## 6     94     59     66
```

```
#Nur noch Daten von 2019
Ind_NYC_2019<-data.frame("neighbourhood_group2"= Ind_NYC$Geographic.Identifier, "Indicator"=Ind_NYC$Indi
head(Ind_NYC_2019)
```

```
##   neighbourhood_group2                    Indicator Incidents
## 1      Staten Island 3 Resolved Consumer Complaints        14
## 2      Staten Island 2 Resolved Consumer Complaints        26
## 3      Staten Island 1 Resolved Consumer Complaints        34
## 4            Queens 14 Resolved Consumer Complaints        25
## 5            Queens 13 Resolved Consumer Complaints        40
## 6            Queens 12 Resolved Consumer Complaints        66
```

```
Ind_NYC_2019_cleaned<-Ind_NYC_2019


#Nummern aus den Distrikten entfernen
Ind_NYC_2019_cleaned$neighbourhood_group <-gsub("[0-9]","", Ind_NYC_2019_cleaned$neighbourhood_group2 )
#Leerzeichen aus den Distrikten entfernen
Ind_NYC_2019_cleaned$neighbourhood_group <-gsub(" ","", Ind_NYC_2019_cleaned$neighbourhood_group )
#Alle distrikte klein schreiben
Ind_NYC_2019_cleaned$neighbourhood_group<-tolower(Ind_NYC_2019_cleaned$neighbourhood_group)
# Neighbourhood Group als Faktor
Ind_NYC_2019_cleaned$neighbourhood_group<-factor(Ind_NYC_2019_cleaned$neighbourhood_group)

#Überblick
head(Ind_NYC_2019_cleaned$Incidents)
```

```
## [1] 14 26 34 25 40 66
```

```
head(Ind_NYC_2019_cleaned$neighbourhood_group)
```

```
## [1] statenisland statenisland statenisland queens       queens
## [6] queens
## Levels:  bronx brooklyn manhattan queens statenisland
```

```
summary(Ind_NYC_2019_cleaned)
```

```
##  neighbourhood_group2
##          : 177
##  Bronx 1 :  35
##  Bronx 10:  35
##  Bronx 11:  35
##  Bronx 2 :  35
##  Bronx 3 :  35
##  (Other) :3307
##                                                              Indicator
##                                                                  : 177
##  Average Response Time to crimes in progress - Critical (minutes):  77
##  Burglary                                                        :  77
##  Crime related to domestic violence - Felonious assault          :  77
##  Crime related to domestic violence - Murder                     :  77
##  Crime related to domestic violence - Rape                       :  77
##  (Other)                                                         :3097
##    Incidents          neighbourhood_group
##  Min.   :     0.0               :1633
##  1st Qu.:    12.6   bronx       : 424
##  Median :    85.6   brooklyn    : 616
##  Mean   :  2319.2   manhattan   : 400
##  3rd Qu.:   322.8   queens      : 480
##  Max.   :424490.0   statenisland: 106
##  NA's   :1181
```

```
summary(Ind_NYC_2019_cleaned$Indicator)
```

```
##
##                                                                                         177
##                                                                     Air complaints received
##                                                                                          59
##                                                                Asbestos complaints received
##                                                                                          59
##                                                                    Average Daily Attendance
##                                                                                          32
##                                                         Average expenditure per student ($)
##                                                                                          32
##                            Average Response Time to crimes in progress - Critical (minutes)
##                                                                                          77
##            Average response time to life-threatening medical emergencies by ambulance units
##                                                                                           5
##                 Average response time to life-threatening medical emergencies by fire units
##                                                                                           5
##                                                   Average response time to structural fires
##                                                                                           5
##                                                                                    Burglary
##                                                                                          77
##                Children in the public schools who have completed required immunizations (%)
##                                                                                          32
##        Citywide acceptability rating for the cleanliness of small parks and playgrounds (%)
##                                                                                          59
## Citywide acceptability rating for the overall condition of small parks and playgrounds (%)
##                                                                                          59
##                                                                    Civilian fire fatalities
##                                                                                          59
##                                      Crime related to domestic violence - Felonious assault
##                                                                                          77
##                                                 Crime related to domestic violence - Murder
##                                                                                          77
##                                                   Crime related to domestic violence - Rape
##                                                                                          77
##                               Curbside and containerized mixed paper recycled tons per day 
##                                                                                         59 
##                                           Curbside and Containerized Recycled Tons Per Day 
##                                                                                         59 
##                                        Curbside and Containerized Recycling Diversion Rate 
##                                                                                         59 
##                                               Deaths from unintentional drug overdose (CY) 
##                                                                                         59 
##                                                       Domestic Violence Related Radio Runs 
##                                                                                         77 
##                                                                          Felonious assault 
##                                                                                         77 
##                                                                              Forcible rape 
##                                                                                         77 
##                                                                              Grand larceny 
##                                                                                         77 
##                                                                         Grand larceny auto 
##                                                                                         77 
##                                                       Hate Crime Related Felonious Assault 
##                                                                                         77 
##                                                                  Hate Crime Related Murder 
##                                                                                         77 
##                                                                        Hate Crimes (total) 
##                                                                                         77 
##                                                                    Intentionally set fires 
##                                                                                         59 
##                                                                         Major felony crime 
##                                                                                         77 
##                                                       Medical Emergencies (fire unit only) 
##                                                                                         59 
##                                                      Murder and non-negligent manslaughter 
##                                                                                         77 
##                          New Cases Requiring Environmental Intervention For Lead Poisoning 
##                                                                                         59 
##                                                                  Noise complaints received 
##                                                                                         59 
##                                                                        Nonstructural Fires 
##                                                                                         59 
##                                       Number of Priority A (emergency) complaints received 
##                                                                                         59 
##                                    Number of Priority B (nonemergency) complaints received 
##                                                                                         59 
##                                                          Persons receiving Cash Assistance 
##                                                                                         59 
##                                                            Persons receiving SNAP benefits 
##                                                                                         59 
##                                                           Private transfer station permits 
##                                                                                         59 
##                                                          Public Health Insurance enrollees 
##                                                                                         59 
##                                                              Recycling tons per truckshift 
##                                                                                         59 
##                                               Refuse Collected for Disposal (tons per day) 
##                                                                                         59 
##                                                                  Refuse tons per truckshift
##                                                                                          59
##                                                                Resolved Consumer Complaints
##                                                                                          59
##                                                   Restaurants scoring an â\200\230Aâ\200\231 grade (
##                                                                                          59
##                                                                                     Robbery
##                                                                                          77
##                                    School Buildings in Good or Fair to Good Condition (%)
##                                                                                          32
##                                                       Sidewalks rated acceptably clean (%)
##                                                                                          59
##                                                                Sidewalks rated filthy (%)
##                                                                                          59
##                                      Streets maintained with a pavement rating of Good (%)
##                                                                                          59
##                                                         Streets rated acceptably clean (%)
##                                                                                          59
##                                                                  Streets rated filthy (%)
##                                                                                          59
##                                                                           Structural Fires
##                                                                                          59
##       Students in grades 3 to 8 meeting or exceeding standards - English Language Arts (%)
##                                                                                          32
##                        Students in grades 3 to 8 meeting or exceeding standards - Math (%)
##                                                                                          32
##                 Students in schools that exceed capacity (%)   - Elementary/middle schools
##                                                                                          32
##                                                             Tons of refuse collected (000)
##                                                                                          59
##                                                               Total housing starts (units)
##                                                                                          59
##                                                                Total Segment 1-8 Incidents
##                                                                                           5
##                                                                          Water main breaks
##                                                                                          59
```

```
levels(Ind_NYC_2019_cleaned$neighbourhood_group)
```

```
## [1] ""             "bronx"        "brooklyn"     "manhattan"
## [5] "queens"       "statenisland"
```

```
# Summe der Incidents pro Distrikt und Indikator
Summary_Ind_NYC_2019<-Ind_NYC_2019_cleaned %>%
  group_by(neighbourhood_group=Ind_NYC_2019_cleaned$neighbourhood_group,Indicator) %>%
  summarise(Observations=sum(Incidents,na.rm = TRUE))
summary(Summary_Ind_NYC_2019)
```

```
##    neighbourhood_group
##              :24
##  bronx       :38
##  brooklyn    :38
##  manhattan   :37
##  queens      :38
##  statenisland:38
```

```
##
##                                                                                    Indicator
##  Air complaints received                                                             :  5
##  Asbestos complaints received                                                        :  5
##  Average response time to life-threatening medical emergencies by ambulance units    :  5
##  Average response time to life-threatening medical emergencies by fire units         :  5
##  Average response time to structural fires                                           :  5
##  Citywide acceptability rating for the cleanliness of small parks and playgrounds (%):  5
##  (Other)                                                                             :183
##   Observations
##  Min.   :     0
##  1st Qu.:     6
##  Median :   273
##  Mean   : 26981
##  3rd Qu.:  2914
##  Max.   :556596
##
```

```
# Angaben ohne Distrikt werte entfernen
Summary_Ind_NYC_2019<-filter(Summary_Ind_NYC_2019,neighbourhood_group != "")
summary(Summary_Ind_NYC_2019)
```

```
##    neighbourhood_group
##              : 0
##  bronx       :38
##  brooklyn    :38
##  manhattan   :37
##  queens      :38
##  statenisland:38
##
##                                                                                    Indicator
##  Air complaints received                                                             :  5
##  Asbestos complaints received                                                        :  5
##  Average response time to life-threatening medical emergencies by ambulance units    :  5
##  Average response time to life-threatening medical emergencies by fire units         :  5
##  Average response time to structural fires                                           :  5
##  Citywide acceptability rating for the cleanliness of small parks and playgrounds (%):  5
##  (Other)                                                                             :159
##   Observations
##  Min.   :     0.0
##  1st Qu.:     7.2
##  Median :   273.0
##  Mean   : 29370.4
##  3rd Qu.:  2617.5
##  Max.   :556596.0
##
```

```
head(Summary_Ind_NYC_2019)
```

```
## # A tibble: 6 x 3
## # Groups:   neighbourhood_group [1]
##   neighbourhood_gro~ Indicator                                 Observations
##   <fct>              <fct>                                            <dbl>
## 1 bronx              Air complaints received                         536
## 2 bronx              Asbestos complaints received                    212
## 3 bronx              Average response time to life-threatenin~         7.44
## 4 bronx              Average response time to life-threatenin~         5.13
## 5 bronx              Average response time to structural fires         4.36
## 6 bronx              Citywide acceptability rating for the cl~      1137.
```

```
# Indikator verschachteln
NYC_nest<-Summary_Ind_NYC_2019 %>%
  nest(Indicator=c(Indicator, Observations))
head(NYC_nest)
```

```
## # A tibble: 5 x 2
##   neighbourhood_group data
##   <fct>               <list>
## 1 bronx               <tibble [38 x 2]>
## 2 brooklyn            <tibble [38 x 2]>
## 3 manhattan           <tibble [37 x 2]>
## 4 queens              <tibble [38 x 2]>
## 5 statenisland        <tibble [38 x 2]>
```

```
#Join both datasets
NYC<-left_join(AB_NYC_available,NYC_nest, by="neighbourhood_group")
```

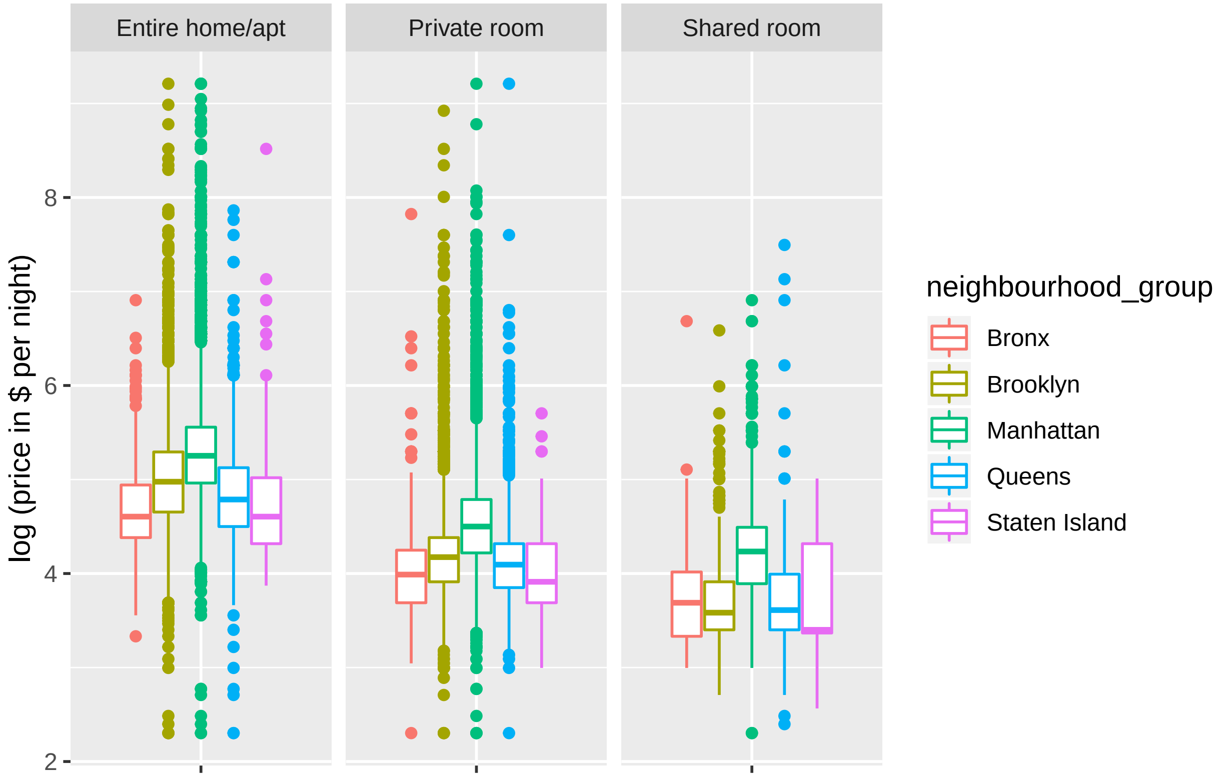
```
## Warning: Column `neighbourhood_group` joining factors with different
## levels, coercing to character vector
```

```
# Neighbourhood Group als Faktor
NYC$neighbourhood_group<-factor(NYC$neighbourhood_group)
```

Data visualisation

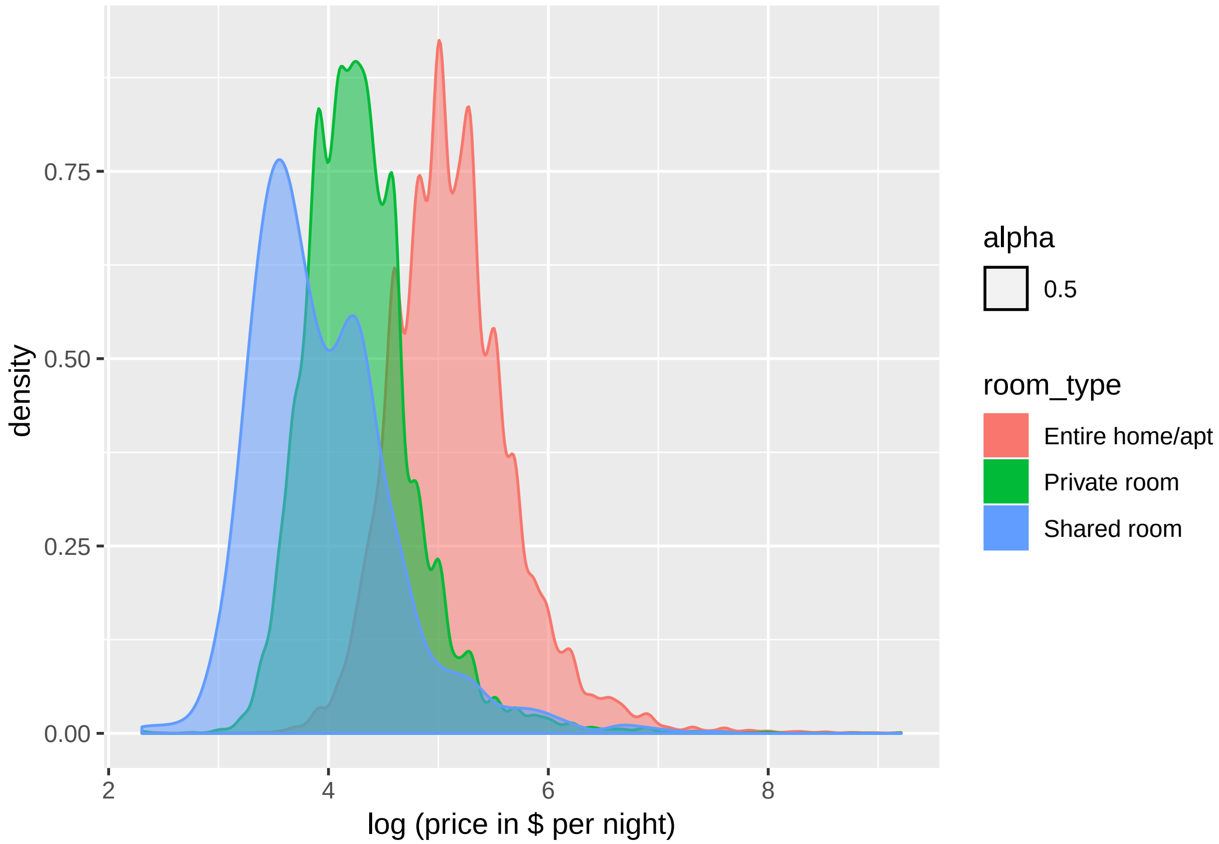
Distribution of prices by room types and neighbourhood

```
ggplot(data = AB_NYC,
       mapping = aes(y = price_log,
                     x = "",
                     group = neighbourhood_group,
                     colour = neighbourhood_group)) +
  geom_boxplot() +
  facet_wrap(. ~ room_type)+
  xlab("")+
  ylab("log (price in $ per night)")
```



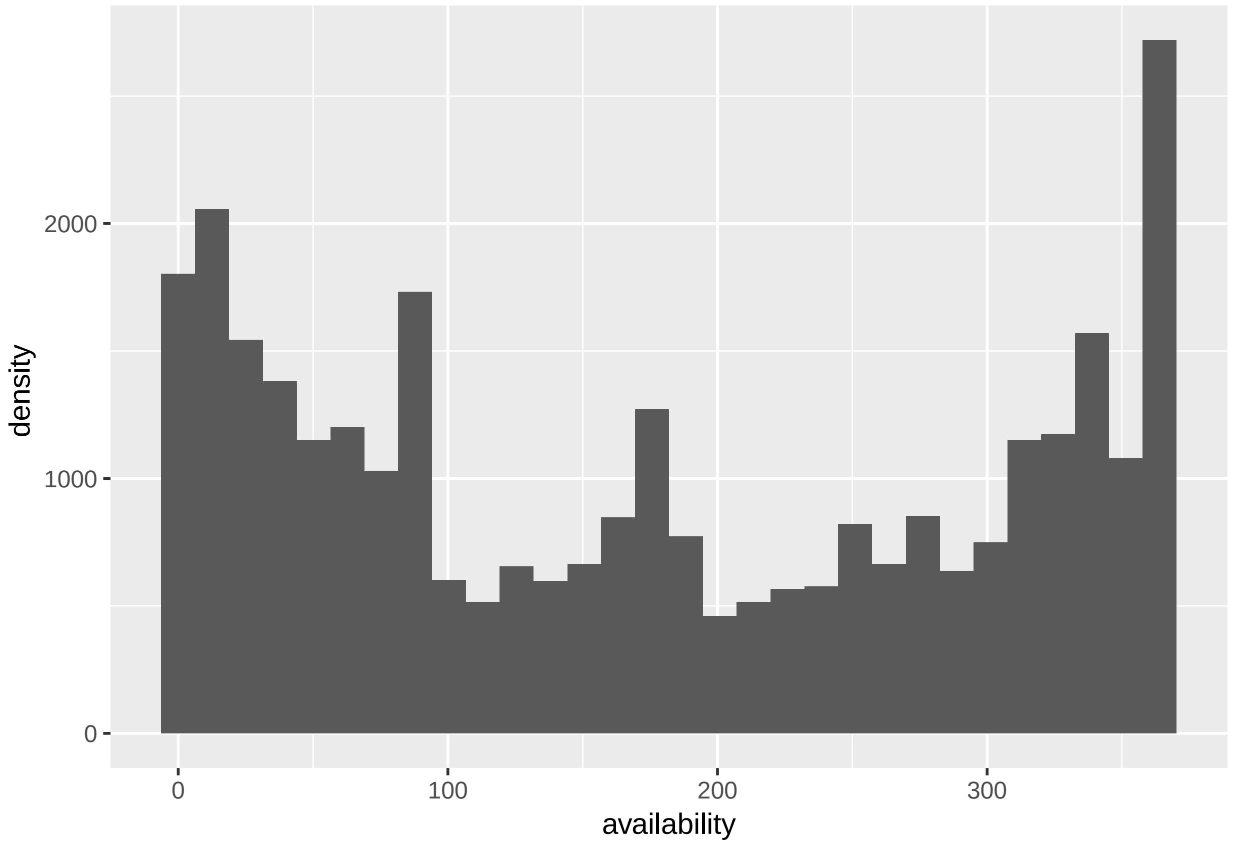
?

```
## price distribution
ggplot(data = AB_NYC,
       mapping = aes(x = price_log,
                     group = room_type,
                     colour = room_type,
                     fill = room_type,
                     alpha = 0.5)) +
  geom_density() +
  xlab("log (price in $ per night)")+
  ylab("density ")
```



```
## availability distribution without 0
ggplot(data = AB_NYC_available,
       mapping = aes(x = availability_365)) +
  geom_histogram() +
  xlab("availability")+
  ylab("density ")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Interactive map with the leaflet package

```
df_exp<-filter(NYC,price == max(price))
df_cheap<-filter(NYC,price == min(price))
```