airbnb in New York City

Carole Mattmann und Jonas Zuercher

13 Februar 2020

Included packages:

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

Introduction

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

```
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 157
##  $ 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 1
##  $ 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

Following changes have been made to the dataset:

remove price 0

remove all listings with price 0

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

Create subsets for room type

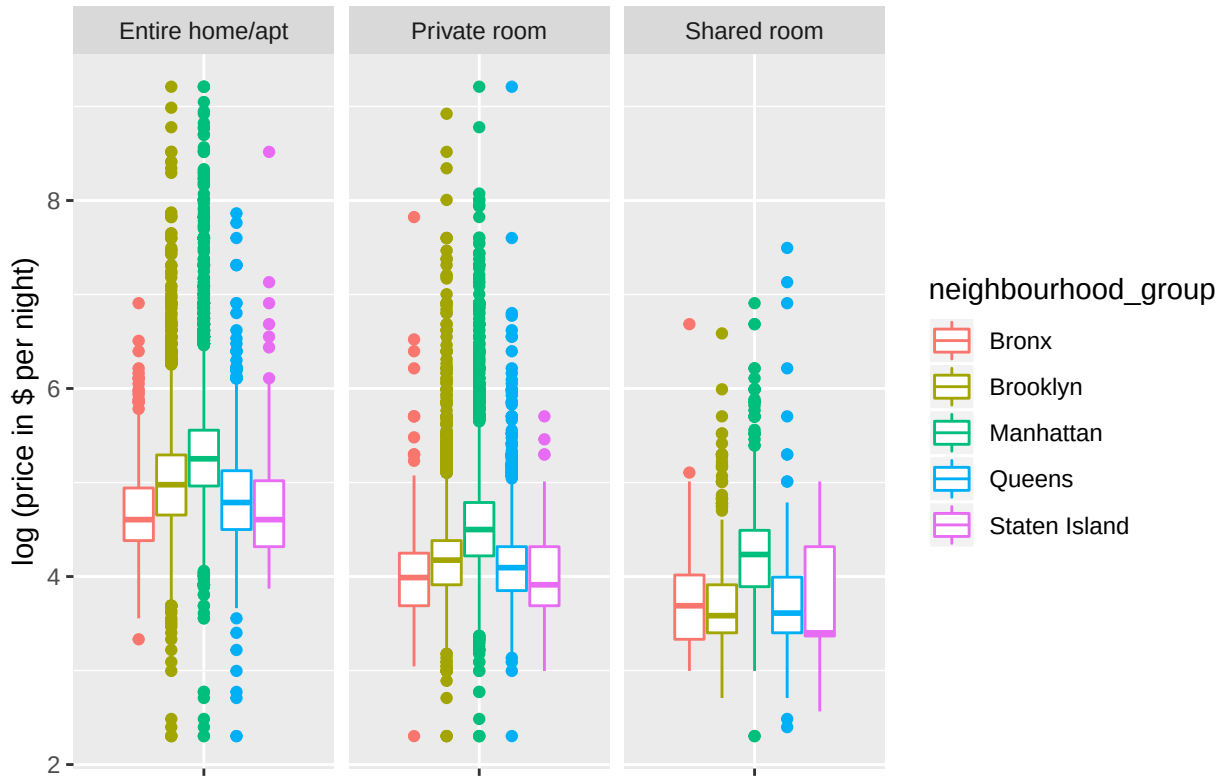
```
AB_NYC_entirehome <-AB_NYC[AB_NYC$room_type == "Entire home/apt",]

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

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

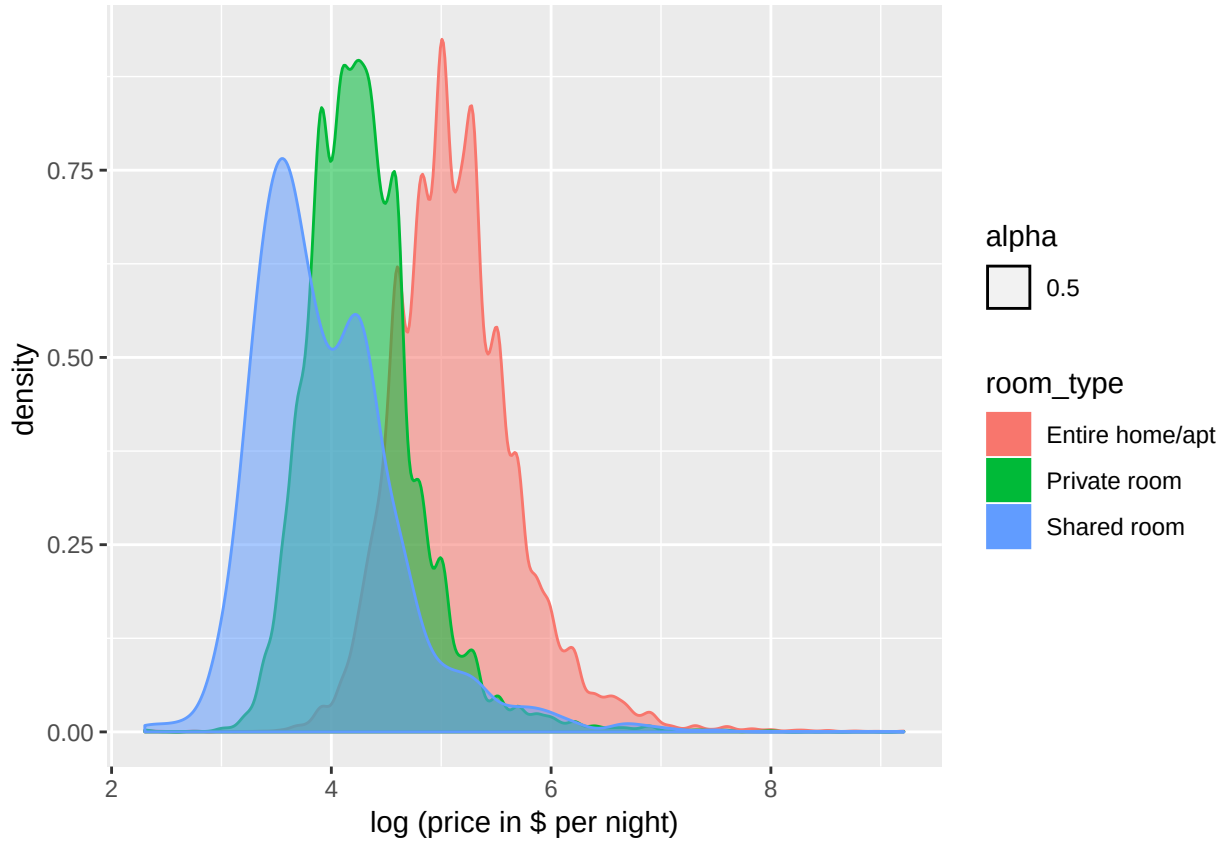
Data visualisation

```
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)")
```



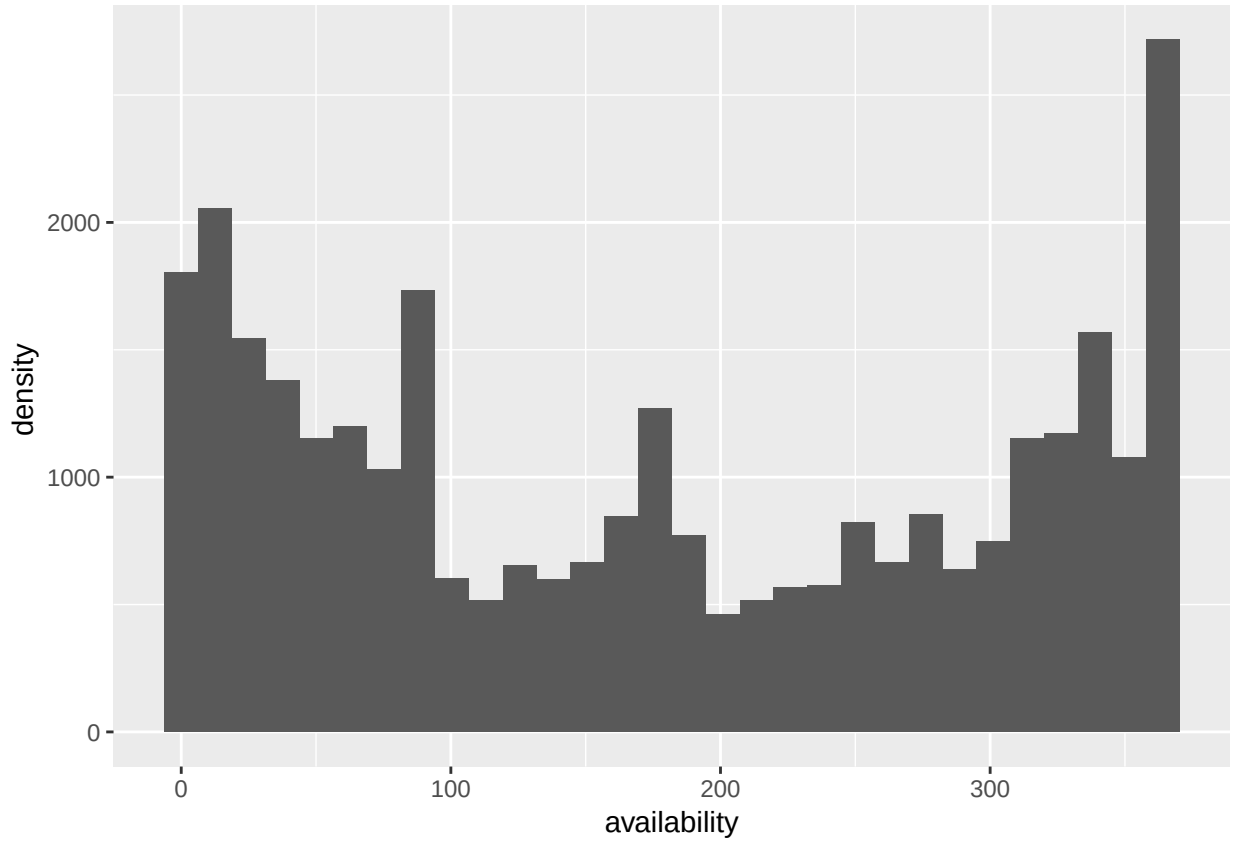
?

```
## 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`.
```



Possible models to calculate the price of an airbnb

```
##simple linear models

lm.hood <- lm (data=AB_NYC, price_log~neighbourhood_group)
summary(lm.hood)
```

```
##
## Call:
## lm(formula = price_log ~ neighbourhood_group, data = AB_NYC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6969 -0.4592 -0.0227  0.4102  4.8391
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.24403    0.01971 215.378  < 2e-16 ***
## neighbourhood_groupBrooklyn     0.32247    0.02023  15.939  < 2e-16 ***
## neighbourhood_groupManhattan    0.75544    0.02019  37.408  < 2e-16 ***
## neighbourhood_groupQueens       0.12718    0.02152   5.911 3.43e-09 ***
## neighbourhood_groupStaten Island  0.12797    0.03903   3.279  0.00104 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6506 on 48879 degrees of freedom
```

```
## Multiple R-squared:  0.1319, Adjusted R-squared:  0.1319
## F-statistic:  1857 on 4 and 48879 DF,  p-value: < 2.2e-16
```

```
lm.type <- lm (data=AB_NYC, price_log~room_type)
summary(lm.type)
```

```
##
## Call:
## lm(formula = price_log ~ room_type, data = AB_NYC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8383 -0.3642 -0.0475  0.2884  4.9143
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.140924   0.003433 1497.38   <2e-16 ***
## room_typePrivate room -0.844881   0.005021 -168.28   <2e-16 ***
## room_typeShared room  -1.188148   0.016444  -72.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5472 on 48881 degrees of freedom
## Multiple R-squared:  0.3857, Adjusted R-squared:  0.3857
## F-statistic: 1.535e+04 on 2 and 48881 DF,  p-value: < 2.2e-16
```

```
lm.dist <- lm (data=AB_NYC, price_log~dist.timessquare)
summary(lm.dist)
```

```
##
## Call:
## lm(formula = price_log ~ dist.timessquare, data = AB_NYC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8355 -0.4611 -0.0426  0.3829  4.5152
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.160e+00  5.506e-03  937.28   <2e-16 ***
## dist.timessquare -6.074e-05  6.549e-07  -92.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6439 on 48882 degrees of freedom
## Multiple R-squared:  0.1496, Adjusted R-squared:  0.1496
## F-statistic:  8601 on 1 and 48882 DF,  p-value: < 2.2e-16
```

```
#distance and room type on price (with interaction)

lm.dist.type.interact <- lm (data=AB_NYC, price~dist.timessquare*room_type)
summary(lm.dist.type.interact)
```

```
##
## Call:
## lm(formula = price ~ dist.timessquare * room_type, data = AB_NYC)
```

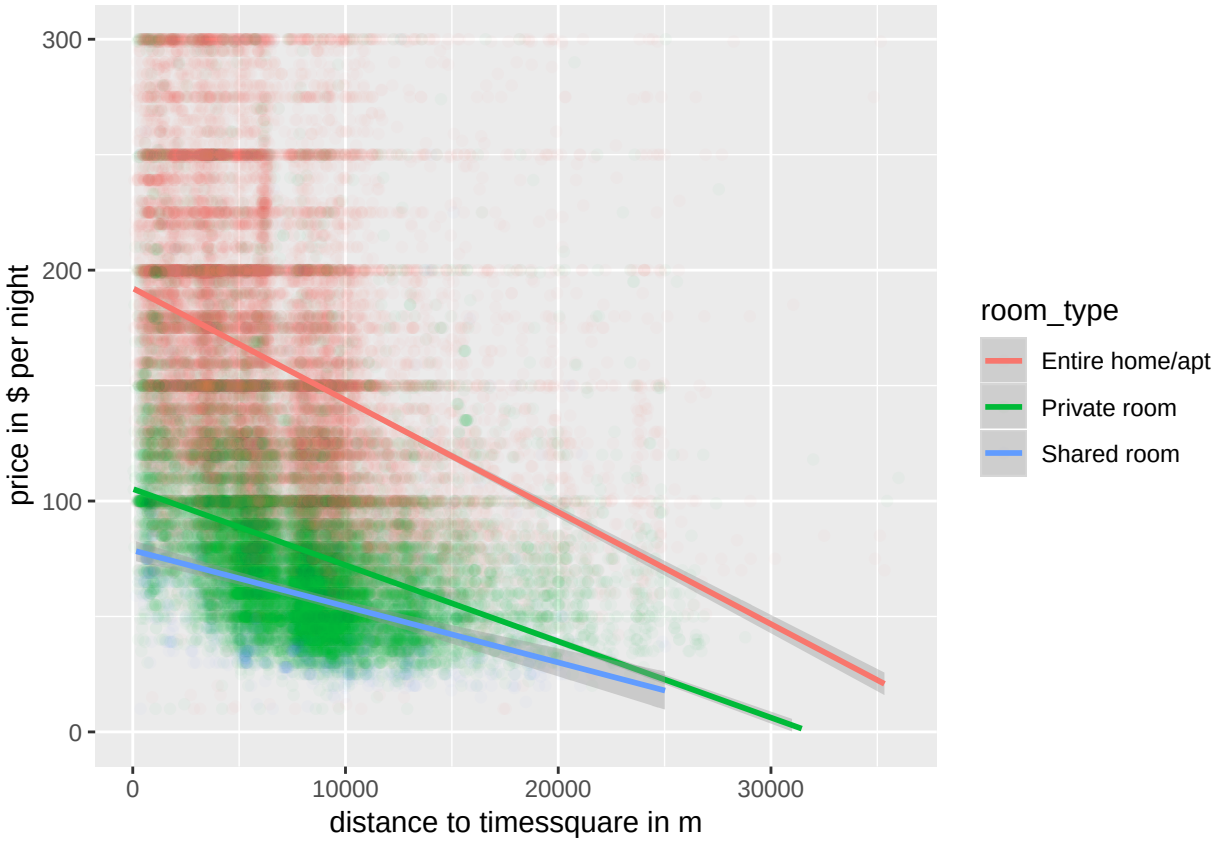
```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -246.7  -55.6  -26.5    7.0 9900.1
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             2.700e+02  2.544e+00 106.158
## dist.timessquare                       -9.119e-03  3.282e-04 -27.783
## room_typePrivate room                  -1.412e+02  4.105e+00 -34.391
## room_typeShared room                   -1.726e+02  1.285e+01 -13.430
## dist.timessquare:room_typePrivate room  4.209e-03  4.844e-04   8.689
## dist.timessquare:room_typeShared room   5.698e-03  1.375e-03   4.143
##                                         Pr(>|t|)
## (Intercept)                              < 2e-16 ***
## dist.timessquare                         < 2e-16 ***
## room_typePrivate room                    < 2e-16 ***
## room_typeShared room                     < 2e-16 ***
## dist.timessquare:room_typePrivate room   < 2e-16 ***
## dist.timessquare:room_typeShared room   3.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 229.9 on 48878 degrees of freedom
## Multiple R-squared:  0.08374,    Adjusted R-squared:  0.08364
## F-statistic: 893.4 on 5 and 48878 DF,  p-value: < 2.2e-16
```

```
ggplot(data = AB_NYC,
       mapping = aes(y = price,
                     x = dist.timessquare,
                     colour = room_type,
                     group = room_type)) +
  geom_point(alpha = 0.03) +
  xlab("distance to timessquare in m")+
  ylab("price in $ per night")+
  ylim(0,300)+
  geom_smooth(method="lm")
```

```
## Warning: Removed 3357 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3357 rows containing missing values (geom_point).
```

```
## Warning: Removed 10 rows containing missing values (geom_smooth).
```




```
#multiple linear regression

lm.full <- lm (data=AB_NYC, price_log~room_type+neighbourhood_group+minimum_nights+number_of_reviews+ca
summary(lm.full)
```

```
##
## Call:
## lm(formula = price_log ~ room_type + neighbourhood_group + minimum_nights +
##     number_of_reviews + calculated_host_listings_count + availability_365 +
##     dist.timessquare, data = AB_NYC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0286 -0.3116 -0.0552  0.2341  5.2323
##
## Coefficients:
##                                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                     5.124e+00  1.778e-02  288.163  < 2e-16
## room_typePrivate room          -7.633e-01  4.646e-03 -164.292  < 2e-16
## room_typeShared room           -1.154e+00  1.493e-02  -77.329  < 2e-16
## neighbourhood_groupBrooklyn     1.165e-01  1.568e-02    7.434 1.07e-13
## neighbourhood_groupManhattan    2.687e-01  1.661e-02   16.177  < 2e-16
## neighbourhood_groupQueens       2.533e-02  1.646e-02    1.539    0.124
## neighbourhood_groupStaten Island  2.173e-01  2.997e-02    7.250 4.22e-13
## minimum_nights                 -2.007e-03  1.117e-04  -17.974  < 2e-16
## number_of_reviews              -7.830e-04  5.158e-05  -15.182  < 2e-16
```

```
## calculated_host_listings_count   -8.389e-05  7.151e-05   -1.173    0.241
## availability_365                  7.931e-04  1.821e-05   43.546  < 2e-16
## dist.timessquare                 -3.486e-05  6.877e-07  -50.689  < 2e-16
##
## (Intercept)                      ***
## room_typePrivate room            ***
## room_typeShared room             ***
## neighbourhood_groupBrooklyn      ***
## neighbourhood_groupManhattan     ***
## neighbourhood_groupQueens
## neighbourhood_groupStaten Island ***
## minimum_nights                   ***
## number_of_reviews                ***
## calculated_host_listings_count
## availability_365                 ***
## dist.timessquare                 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4942 on 48872 degrees of freedom
## Multiple R-squared:  0.4992, Adjusted R-squared:  0.4991
## F-statistic:  4429 on 11 and 48872 DF,  p-value: < 2.2e-16
```

```
lm.empty <- lm (data=AB_NYC, price_log~NULL)
add1(lm.empty,scope=lm.full)
```

```
## Single term additions
##
## Model:
## price_log ~ NULL
##                                Df Sum of Sq   RSS    AIC
## <none>                                      23831 -35119
## room_type                       2    9192.4 14639 -58936
## neighbourhood_group             4    3144.1 20687 -42027
## minimum_nights                  1      26.3 23805 -35171
## number_of_reviews               1      42.7 23789 -35204
## calculated_host_listings_count  1     419.4 23412 -35985
## availability_365                1     232.8 23599 -35597
## dist.timessquare                1    3565.7 20266 -43040
```

```
#choose value with smallest RSS

lm.1 <- update(lm.empty,.~.+room_type)
add1(lm.1,scope=lm.full)
```

```
## Single term additions
##
## Model:
## price_log ~ room_type
##                                Df Sum of Sq   RSS    AIC
## <none>                                      14639 -58936
## neighbourhood_group             4   1633.63 13005 -64713
## minimum_nights                  1      3.69 14635 -58947
## number_of_reviews               1     34.47 14604 -59050
## calculated_host_listings_count  1     98.69 14540 -59265
```

```
## availability_365                1    275.13 14364 -59862
## dist.timessquare                1   1905.94 12733 -65753
```

```
lm.2 <- update(lm.1,.~.+dist.timessquare)
add1(lm.2,scope=lm.full)
```

```
## Single term additions
##
## Model:
## price_log ~ room_type + dist.timessquare
##                                Df Sum of Sq   RSS    AIC
## <none>                                      12733 -65753
## neighbourhood_group             4    276.66 12456 -66819
## minimum_nights                  1     22.26 12711 -65837
## number_of_reviews               1     11.12 12722 -65794
## calculated_host_listings_count  1     31.20 12702 -65871
## availability_365                1    417.06 12316 -67379
```

```
lm.3 <- update(lm.2,.~.+availability_365)
add1(lm.3,scope=lm.full)
```

```
## Single term additions
##
## Model:
## price_log ~ room_type + dist.timessquare + availability_365
##                                Df Sum of Sq   RSS    AIC
## <none>                                      12316 -67379
## neighbourhood_group             4   258.228 12058 -68407
## minimum_nights                  1    62.412 12254 -67625
## number_of_reviews               1    47.601 12268 -67566
## calculated_host_listings_count  1     0.538 12315 -67379
```

```
lm.4 <- update(lm.3,.~.+neighbourhood_group)
add1(lm.4,scope=lm.full)
```

```
## Single term additions
##
## Model:
## price_log ~ room_type + dist.timessquare + availability_365 +
##     neighbourhood_group
##                                Df Sum of Sq   RSS    AIC
## <none>                                      12058 -68407
## minimum_nights                  1    67.421 11990 -68679
## number_of_reviews               1    43.380 12014 -68581
## calculated_host_listings_count  1     0.229 12058 -68406
```

```
lm.5 <- update(lm.4,.~.+minimum_nights)
add1(lm.5,scope=lm.full)
```

```
## Single term additions
##
## Model:
## price_log ~ room_type + dist.timessquare + availability_365 +
##     neighbourhood_group + minimum_nights
##                                Df Sum of Sq   RSS    AIC
## <none>                                      11990 -68679
## number_of_reviews               1    55.981 11934 -68906
```

```
## calculated_host_listings_count  1     0.036 11990 -68677
```

```
summary(lm.5)
```

```
##
## Call:
## lm(formula = price_log ~ room_type + dist.timessquare + availability_365 +
##     neighbourhood_group + minimum_nights, data = AB_NYC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0121 -0.3120 -0.0582  0.2311  5.2284
##
## Coefficients:
##                                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                     5.112e+00  1.781e-02  287.116  < 2e-16
## room_typePrivate room          -7.628e-01  4.641e-03 -164.355  < 2e-16
## room_typeShared room           -1.145e+00  1.494e-02  -76.671  < 2e-16
## dist.timessquare               -3.502e-05  6.888e-07  -50.838  < 2e-16
## availability_365                7.388e-04  1.744e-05   42.355  < 2e-16
## neighbourhood_groupBrooklyn     1.138e-01  1.571e-02    7.243 4.45e-13
## neighbourhood_groupManhattan    2.671e-01  1.664e-02   16.054  < 2e-16
## neighbourhood_groupQueens       2.231e-02  1.650e-02    1.353    0.176
## neighbourhood_groupStaten Island  2.165e-01  3.004e-02    7.207 5.81e-13
## minimum_nights                 -1.841e-03  1.111e-04  -16.578  < 2e-16
##
## (Intercept)                     ***
## room_typePrivate room           ***
## room_typeShared room            ***
## dist.timessquare                ***
## availability_365                ***
## neighbourhood_groupBrooklyn     ***
## neighbourhood_groupManhattan    ***
## neighbourhood_groupQueens
## neighbourhood_groupStaten Island ***
## minimum_nights                  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4953 on 48874 degrees of freedom
## Multiple R-squared:  0.4969, Adjusted R-squared:  0.4968
## F-statistic:  5363 on 9 and 48874 DF,  p-value: < 2.2e-16
```

```
lm.full2 <- lm (data=AB_NYC_entirehome, price_log~neighbourhood_group+minimum_nights+number_of_reviews+
summary(lm.full2)
```

```
##
## Call:
## lm(formula = price_log ~ neighbourhood_group + minimum_nights +
##     number_of_reviews + calculated_host_listings_count + availability_365 +
##     dist.timessquare, data = AB_NYC_entirehome)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0512 -0.3263 -0.0568  0.2520  4.0934
```

```
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.052e+00  3.009e-02 167.886  < 2e-16
## neighbourhood_groupBrooklyn     2.121e-01  2.732e-02   7.763 8.63e-15
## neighbourhood_groupManhattan    3.036e-01  2.869e-02  10.579  < 2e-16
## neighbourhood_groupQueens       6.189e-02  2.879e-02   2.150   0.0316
## neighbourhood_groupStaten Island  3.027e-01  4.728e-02   6.401 1.57e-10
## minimum_nights                 -2.221e-03  1.448e-04 -15.340  < 2e-16
## number_of_reviews              -1.250e-03  7.931e-05 -15.765  < 2e-16
## calculated_host_listings_count  -4.125e-05  7.804e-05  -0.529   0.5971
## availability_365                1.012e-03  2.722e-05  37.169  < 2e-16
## dist.timessquare               -3.462e-05  1.051e-06 -32.941  < 2e-16
##
## (Intercept)                     ***
## neighbourhood_groupBrooklyn     ***
## neighbourhood_groupManhattan    ***
## neighbourhood_groupQueens       *
## neighbourhood_groupStaten Island ***
## minimum_nights                  ***
## number_of_reviews               ***
## calculated_host_listings_count
## availability_365                ***
## dist.timessquare                ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5127 on 25397 degrees of freedom
## Multiple R-squared:  0.1837, Adjusted R-squared:  0.1834
## F-statistic:   635 on 9 and 25397 DF,  p-value: < 2.2e-16
```

```
lm.full3 <- lm (data=AB_NYC_privateroom, price_log~neighbourhood_group+minimum_nights+number_of_reviews
summary(lm.full3)
```

```
##
## Call:
## lm(formula = price_log ~ neighbourhood_group + minimum_nights +
##     number_of_reviews + calculated_host_listings_count + availability_365 +
##     dist.timessquare, data = AB_NYC_privateroom)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3265 -0.2871 -0.0494  0.2111  5.1882
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.411e+00  2.187e-02 201.700  < 2e-16
## neighbourhood_groupBrooklyn     5.963e-02  1.905e-02   3.131  0.00175
## neighbourhood_groupManhattan    2.626e-01  2.028e-02  12.951  < 2e-16
## neighbourhood_groupQueens       1.918e-02  1.990e-02   0.964  0.33511
## neighbourhood_groupStaten Island  1.619e-01  3.850e-02   4.206 2.61e-05
## minimum_nights                 -1.917e-03  1.918e-04  -9.997  < 2e-16
## number_of_reviews              -5.155e-04  6.660e-05  -7.740 1.03e-14
## calculated_host_listings_count  -2.645e-03  3.102e-04  -8.527  < 2e-16
## availability_365                7.039e-04  2.469e-05  28.515  < 2e-16
```

```
## dist.timessquare                 -3.610e-05  9.086e-07 -39.734  < 2e-16
##
## (Intercept)                      ***
## neighbourhood_groupBrooklyn      **
## neighbourhood_groupManhattan     ***
## neighbourhood_groupQueens
## neighbourhood_groupStaten Island ***
## minimum_nights                   ***
## number_of_reviews                ***
## calculated_host_listings_count   ***
## availability_365                 ***
## dist.timessquare                 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4606 on 22309 degrees of freedom
## Multiple R-squared:  0.2071, Adjusted R-squared:  0.2067
## F-statistic: 647.3 on 9 and 22309 DF,  p-value: < 2.2e-16
```

```
lm.full4 <- lm (data=AB_NYC_sharedroom, price_log~neighbourhood_group+minimum_nights+number_of_reviews+
summary(lm.full4)
```

```
##
## Call:
## lm(formula = price_log ~ neighbourhood_group + minimum_nights +
##     number_of_reviews + calculated_host_listings_count + availability_365 +
##     dist.timessquare, data = AB_NYC_sharedroom)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1783 -0.3299 -0.0873  0.2001  3.5019
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.215e+00  9.455e-02  44.581  < 2e-16
## neighbourhood_groupBrooklyn      1.287e-02  7.947e-02   0.162 0.871370
## neighbourhood_groupManhattan     3.051e-01  8.730e-02   3.495 0.000493
## neighbourhood_groupQueens       -1.883e-02  8.468e-02  -0.222 0.824078
## neighbourhood_groupStaten Island 2.446e-01  2.069e-01   1.182 0.237470
## minimum_nights                  -5.130e-04  5.374e-04  -0.955 0.339986
## number_of_reviews               -1.977e-03  4.969e-04  -3.978 7.39e-05
## calculated_host_listings_count  -2.404e-02  2.922e-03  -8.226 5.19e-16
## availability_365                -8.164e-05  1.194e-04  -0.684 0.494423
## dist.timessquare                -2.896e-05  4.662e-06  -6.212 7.29e-10
##
## (Intercept)                      ***
## neighbourhood_groupBrooklyn
## neighbourhood_groupManhattan     ***
## neighbourhood_groupQueens
## neighbourhood_groupStaten Island
## minimum_nights
## number_of_reviews                ***
## calculated_host_listings_count   ***
## availability_365
## dist.timessquare                 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5703 on 1148 degrees of freedom
## Multiple R-squared:  0.2395, Adjusted R-squared:  0.2336
## F-statistic: 40.18 on 9 and 1148 DF,  p-value: < 2.2e-16
```

Interactive map with the leaflet package