Data Cleaning Code

Zhaoqi Liu

4/17/2020

### Dataframe abNYC

setwd("~/Downloads/STOR 565/STOR 565 Group Project/new-york-city-airbnb-open-data")   
abnbNYC<-read.csv("AB\_NYC\_2019\_with\_cr.csv")  
abNYC<-abnbNYC[,c(1,5:12,14,15,17)] #remove room\_name, host\_id, host\_name, last\_review\_day  
prop\_availability<-abnbNYC$availability\_365/365  
abNYC$prop.availability<-prop\_availability #make availability\_365 a proportion   
last\_review\_date<-abnbNYC$last\_review  
last\_review\_date<-as.Date(last\_review\_date,format="%d/%m/%Y")  
days\_last\_review<-as.numeric(as.Date("2019-07-09",format="%Y-%m-%d")-last\_review\_date)  
abNYC$days\_last\_review<-days\_last\_review #make last review date as a number   
abNYC<-na.omit(abNYC) #38803 observations   
  
#remove listings with 0 available date and last review over 90 days  
library(dplyr)  
prop\_availability<-abNYC$prop.availability  
days\_last\_review<-abNYC$days\_last\_review  
abNYC<-filter(abNYC,prop\_availability!=0 | days\_last\_review<90)  
#28010 obs, 14 predictors

### Dataframe abNYCnum---Make Neighborhood\_group and room\_type as numeric

#### For PCR model (maybe?)

abNYCnum<-abNYC  
abNYCnum$neighbourhood\_group<-as.numeric(abNYC$neighbourhood\_group)  
#"Bronx"=1,"Brooklyn"=2,"Manhattan"=3,"Queens"=4,"Staten Island"=5  
abNYCnum$room\_type<-as.numeric(abNYC$room\_type)  
#"Entire home/apt=1 "Private room"=2 "Shared room"=3

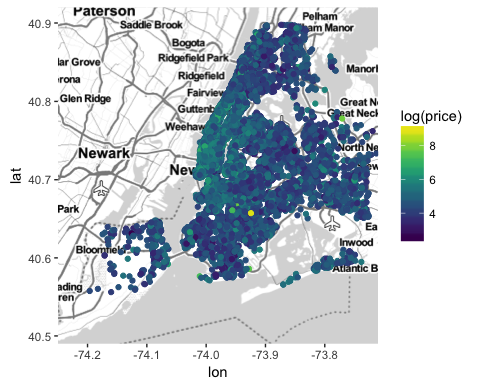
### Split train and test set

#split train and test set   
set.seed(1234)  
train<-sample(nrow(abNYC),size=floor(nrow(abNYC)/4\*3),replace = FALSE)  
train.set<-abNYC[train,] #21007 obs   
test.set<-abNYC[-train,] #7003 obs

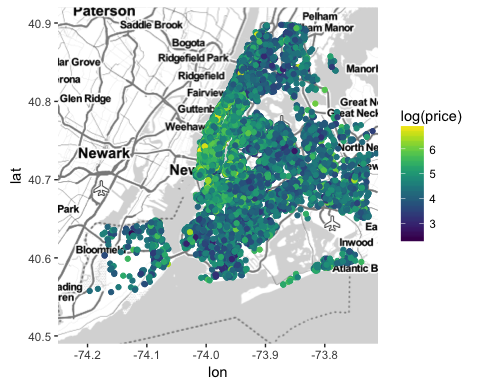
# ggplot for this dataset

library(ggmap)  
nyc <- c(left = -74.25, bottom = 40.49, right = -73.71, top =40.92)  
map<-get\_stamenmap(nyc, zoom = 10, maptype = "toner-lite") %>% ggmap()   
#maptype = c("terrain",  
# "terrain-background", "terrain-labels", "terrain-lines", "toner",  
# "toner-2010", "toner-2011", "toner-background", "toner-hybrid",  
# "toner-labels", "toner-lines", "toner-lite", "watercolor")  
#这些是所有的maptype 都可以选 我觉得“tone-lite” 看起来最简洁

library(ggplot2)  
library(viridis)  
map+geom\_point(aes(x=longitude,y=latitude,col=log(price)),  
 data=abNYC)+  
 scale\_color\_viridis(option = "D")+scale\_size\_area(max\_size = 10)



map+geom\_point(aes(x=longitude,y=latitude,col=log(price)),  
 data=abNYC[abNYC$price<1000,])+  
 scale\_color\_viridis(option = "D")+scale\_size\_area(max\_size = 1)



map+geom\_point(aes(x=longitude,y=latitude,col=log(price)),  
 data=abNYC[abNYC$price<200,])+  
 scale\_color\_viridis(option = "D")+scale\_size\_area(max\_size = 10)

