[Kaggle]

New York City Airbnb Open Data

허지연

INDEX

01 | 수집된 데이터

- 데이터 갯수: 총 48,895개
- 데이터 열: 총 16개 (id, name, host_id, host_name_neiborhood_group, neighborhood, latitude, room_type, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, availability_365)

```
> airbnb <- read.csv("./2021-1학교수업/데이터분석_전처리발표/AB_NYC_2019.csv",stringsAsFactors = FALSE ,na.strings = c(""))
> str(airbnb)
'data.frame':   48895 obs. of  16 variables:
 $ id                            : int  2539 2595 3647 3831 5022 5099 5121 5178 5203 5238 ...
 $ name                          : chr  "Clean & quiet apt home by the park" "Skylit Midtown Castle" "THE VILLAGE OF
 HARLEM....NEW YORK !" "Cozy Entire Floor of Brownstone" ...
 $ host_id                       : int  2787 2845 4632 4869 7192 7322 7356 8967 7490 7549 ...
 $ host_name                     : chr  "John" "Jennifer" "Elisabeth" "LisaRoxanne" ...
 $ neighbourhood_group           : chr  "Brooklyn" "Manhattan" "Manhattan" "Brooklyn" ...
 $ neighbourhood                 : chr  "Kensington" "Midtown" "Harlem" "Clinton Hill" ...
 $ latitude                      : num  40.6 40.8 40.8 40.7 40.8 ...
 $ longitude                     : num  -74 -74 -73.9 -74 -73.9 ...
 $ room_type                     : chr  "Private room" "Entire home/apt" "Private room" "Entire home/apt" ...
 $ 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                   : chr  "2018-10-19" "2019-05-21" NA "2019-07-05" ...
 $ 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 ...
```

01 | 수집된 데이터

```
> summary(airbnb)
       id                name              host_id            host_name         neighbourhood_group
 Min.   :    2539   Length:48895       Min.   :     2438   Length:48895       Length:48895
 1st Qu.: 9471945   Class :character   1st Qu.:  7822033   Class :character   Class :character
 Median :19677284   Mode  :character   Median : 30793816   Mode  :character   Mode  :character
 Mean   :19017143                      Mean   : 67620011
 3rd Qu.:29152178                      3rd Qu.:107434423
 Max.   :36487245                      Max.   :274321313

 neighbourhood         latitude       longitude       room_type             price         minimum_nights
 Length:48895       Min.   :40.50   Min.   :-74.24   Length:48895       Min.   :    0.0   Min.   :   1.00
 Class :character   1st Qu.:40.69   1st Qu.:-73.98   Class :character   1st Qu.:   69.0   1st Qu.:   1.00
 Mode  :character   Median :40.72   Median :-73.96   Mode  :character   Median :  106.0   Median :   3.00
                    Mean   :40.73   Mean   :-73.95                      Mean   :  152.7   Mean   :   7.03
                    3rd Qu.:40.76   3rd Qu.:-73.94                      3rd Qu.:  175.0   3rd Qu.:   5.00
                    Max.   :40.91   Max.   :-73.71                      Max.   :10000.0   Max.   :1250.00

 number_of_reviews last_review        reviews_per_month calculated_host_listings_count availability_365
 Min.   :  0.00    Length:48895       Min.   : 0.010    Min.   :  1.000                 Min.   :  0.0
 1st Qu.:  1.00    Class :character   1st Qu.: 0.190    1st Qu.:  1.000                 1st Qu.:  0.0
 Median :  5.00    Mode  :character   Median : 0.720    Median :  1.000                 Median : 45.0
 Mean   : 23.27                       Mean   : 1.373    Mean   :  7.144                 Mean   :112.8
 3rd Qu.: 24.00                       3rd Qu.: 2.020    3rd Qu.:  2.000                 3rd Qu.:227.0
 Max.   :629.00                       Max.   :58.500    Max.   :327.000                 Max.   :365.0
                                      NA's   :10052
```

적용한 전처리 방법

1) 데이터형 변형
2) 결측치 확인 및 처리
3) 변수 특징 살펴보기 (시각화)
4) 변수 간 상관 관계

1. 데이터형 변형

- 분석에 필요 없을 것 같은 변수들 제거
  - Id, host_id 제거

```
> names_to_delete <- c("id", "host_id")
> airbnb[names_to_delete] <- NULL
```

- Factor 변환
  - host_name
  - neighbourhood_group
  - neighbourhood
  - room_type

```
library(purrr)
names_to_factor <- c("host_name", "neighbourhood_group", "neighbourhood", "room_type")
airbnb[names_to_factor] <- map(airbnb[names_to_factor], as.factor)
```

- 날짜 ymd로 변환

```
install.packages('lubridate')
library(lubridate)
airbnb[c("last_review")]<-ymd(airbnb$last_review)
```

1. 데이터형 변형

```
> str(airbnb)
'data.frame':   48895 obs. of  14 variables:
 $ name                          : chr  "Clean & quiet apt home by the park" "Skylit Midtown Castle" "THE VILLAGE OF HARLEM....NEW YORK !" "Cozy Ent
ire Floor of Brownstone" ...
 $ host_name                     : Factor w/ 11452 levels " Valéria","-TheQueensCornerLot",..: 4996 4790 2912 6209 5928 1937 3548 9648 6879 1234 ...
 $ 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 62 138 14 96 203 36 ...
 $ 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                   : Date, format: "2018-10-19" "2019-05-21" NA "2019-07-05" ...
 $ 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 ...
```

2. 결측치 확인

complete cases: 행에 누락된 데이터가 없는(NA가 존재하지 않는)지를 확인해주는 함수로서 해당 행 전체에 누락된 데이터가 없다면 TRUE값을 반환하고 누락된 데이터가 존재한다면 FALSE를 반환함

- colSums(is.na(airbnb)): 누락된 데이터 확인하기
- airbnb[!complete.cases(airbnb),] : 누락된 데이터 행 추출 **!: 반대를 의미(T/F→F/T)**
- dim(airbnb[!complete.cases(airbnb),]) : 누락된 데이터 행 개수 추출

```
> colSums(is.na(airbnb))
                                id                               name                            host_id
                                 0                                 16                                  0
                         host_name                neighbourhood_group                      neighbourhood
                                21                                  0                                  0
                          latitude                          longitude                          room_type
                                 0                                  0                                  0
                             price                      minimum_nights                  number_of_reviews
                                 0                                  0                                  0
                       last_review                reviews_per_month               calculated_host_listings_count
                             10052                              10052                                  0
                  availability_365
                                 0
```

```
> dim(airbnb[!complete.cases(airbnb),])
[1] 10074    16
```

```
> airbnb[!complete.cases(airbnb),]
       id                                          name host_id       host_name neighbourhood_group       neighbourhood latitude longitude        room_type price minimum_nights number_of_reviews last_review reviews_per_month calculated_host_listings_count availability_365
3    3647                  THE VILLAGE OF HARLEM....NEW YORK !    4632       Elisabeth           Manhattan              Harlem 40.80902 -73.94190     Private room   150               3                 0        <NA>              NA                                 1              365
20   7750                    Huge 2 BR Upper East  Cental Park   17985            Sing           Manhattan       East Harlem 40.79685 -73.94872 Entire home/apt   190               7                 0        <NA>              NA                                 2              249
27   8700 Magnifique Suite au N de Manhattan - vue Cloitres   26394 Claude & Sophie           Manhattan              Inwood 40.86754 -73.92639     Private room    80               4                 0        <NA>              NA                                 1                0
37  11452                           Clean and Quiet in Brooklyn    7355              Vt            Brooklyn Bedford-Stuyvesant 40.68876 -73.94312     Private room    35              60                 0        <NA>              NA                                 1              365
39  11943                             Country space in the city   45445         Harriet            Brooklyn            Flatbush 40.63702 -73.96327     Private room   150               1                 0        <NA>              NA                                 1              365
194 51438                  1 Bedroom in 2 Bdrm Apt- Upper East  236421         Jessica           Manhattan  Upper East Side 40.77333 -73.95199     Private room   130              14                 0        <NA>              NA                                 2                0
205 54466                  Beautiful Uptown Manhattan apartmnt  253385         Douglas           Manhattan              Harlem 40.80234 -73.95603     Private room   200              30                 0        <NA>              NA                                 1              365
261 63588                                                LL3  295128     Carol Gloria              Bronx       Clason Point 40.81309 -73.85514     Private room    90               2                 0        <NA>              NA                                 7              349
266 63913              HOSTING YOUR  SUNNY, SPACIOUS NYC ROOM  312288           Paula           Manhattan              Inwood 40.86648 -73.92630     Private room    75               7                 0        <NA>              NA                                 2              323
268 64015                       Prime East Village 1 Bedroom  146944           David           Manhattan       East Village 40.72807 -73.98594 Entire home/apt   200               3                 0        <NA>              NA                                 1                0
277 65556                   Room in S3rd/Bedford, Williamsburg  320422          Marlon            Brooklyn  Williamsburg 40.71368 -73.96260     Private room    60               3                 0        <NA>              NA                                 1                0
346 89427                                The Brooklyn Waverly  116599            Sahr            Brooklyn      Clinton Hill 40.68613 -73.96536 Entire home/apt   650               5                 0        <NA>              NA                                 3              365
```

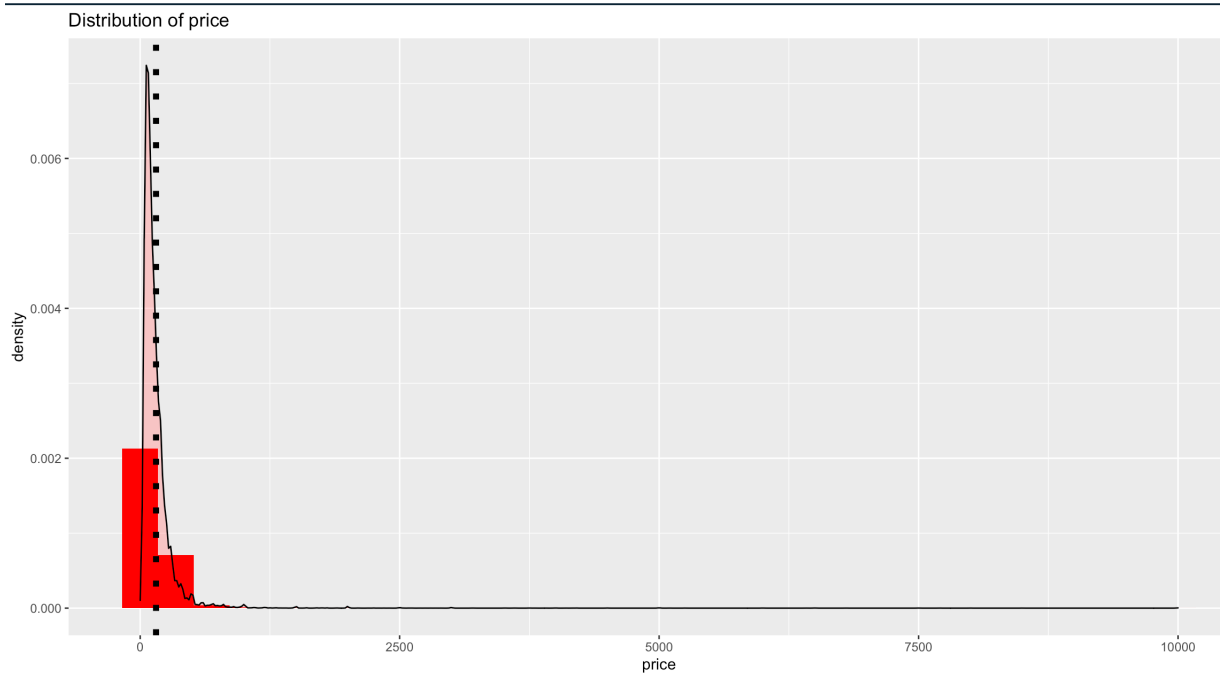
2. 결측치 확인

- airbnb[complete.cases(airbnb),]: 누락된 데이터 제거한 데이터
- 총 48,895 → 총 38,821 (10,074 개 제거됨)

```
> str(airbnb[complete.cases(airbnb),])
'data.frame':   38821 obs. of  14 variables:
 $ name                          : chr  "Clean & quiet apt home by the park" "Skylit Midtown Castle" "Cozy Entire Floor of Brownstone" "Entire Apt:
 Spacious Studio/Loft by central park" ...
 $ host_name                     : Factor w/ 11452 levels " Valéria","-TheQueensCornerLot",..: 4996 4790 6209 5928 1937 3548 9648 6879 1234 6029 ...
 $ neighbourhood_group           : Factor w/ 5 levels "Bronx","Brooklyn",..: 2 3 2 3 3 2 3 3 3 3 ...
 $ neighbourhood                 : Factor w/ 221 levels "Allerton","Arden Heights",..: 109 128 42 62 138 14 96 203 36 203 ...
 $ latitude                      : num  40.6 40.8 40.7 40.8 40.7 ...
 $ longitude                     : num  -74 -74 -74 -73.9 -74 ...
 $ room_type                     : Factor w/ 3 levels "Entire home/apt",..: 2 1 1 1 1 2 2 2 1 1 ...
 $ price                         : int  149 225 89 80 200 60 79 79 150 135 ...
 $ minimum_nights                : int  1 1 1 10 3 45 2 2 1 5 ...
 $ number_of_reviews             : int  9 45 270 9 74 49 430 118 160 53 ...
 $ last_review                   : Date, format: "2018-10-19" "2019-05-21" "2019-07-05" "2018-11-19" ...
 $ reviews_per_month             : num  0.21 0.38 4.64 0.1 0.59 0.4 3.47 0.99 1.33 0.43 ...
 $ calculated_host_listings_count: int  6 2 1 1 1 1 1 1 4 1 ...
 $ availability_365              : int  365 355 194 0 129 0 220 0 188 6 ...
```

3. 변수 특징 살펴보기 (시각화)

- Price 기준 변수 비교 시각화 (가장 중요한 변수)




```r
library(ggplot2)
library(dplyr)
ggplot(airbnb, aes(price))+
  geom_histogram(bins = 30, aes(y = ..density..), fill = "red") +
  geom_density(alpha = 0.2, fill = "red") +
  ggtitle("Distribution of price")+
  theme(axis.title = element_text(), axis.title.x = element_text()) +
  geom_vline(xintercept = round(mean(airbnb$price), 2), size = 2, linetype = 3)
```

X 절편 =price의 mean값

3. 변수 특징 살펴보기 (시각화)

- 한쪽으로 치우쳐 있는 그래프
  - Log 변환:

1. 정규성을 높이고 분석에서 정확한 값을 얻기 위함
2. 데이터 간 편차를 줄여 1) 왜도 2)첨도를 줄일 수 있음
3. 큰 수를 작게 만들 경우 2/ 복잡한 계산을 간편하게 위할 경우 사용된다.


```
ggplot(airbnb, aes(price)) +
  geom_histogram(bins = 30, aes(y = ..density..), fill = "red") +
  geom_density(alpha = 0.2, fill = "purple") +
  ggtitle("Transformed distribution of price") +
  geom_vline(xintercept = round(mean(airbnb$price), 2), size = 2, linetype = 3) +
  scale_x_log10() +
  annotate("text", x = 1800, y = 0.75,label = paste("Mean price = ", paste0(round(mean(airbnb$price), 2), "$")),
           color =  "red", size = 8)
```

3. 변수 특징 살펴보기 (시각화)

- 이웃 지역에 대한 log10 변환을 사용한 neighbourhood_group별 price 히스토그램 밀도 확인
  1. Brooklyn
  2. Manhattan
  3. Queens
  4. Staten Island
  5. The Bronx

```
$ neighbourhood_group           : Factor w/ 5 levels
```

```
> unique(airbnb$neighbourhood_group)
[1] Brooklyn      Manhattan     Queens        Staten Island Bronx
Levels: Bronx Brooklyn Manhattan Queens Staten Island
```

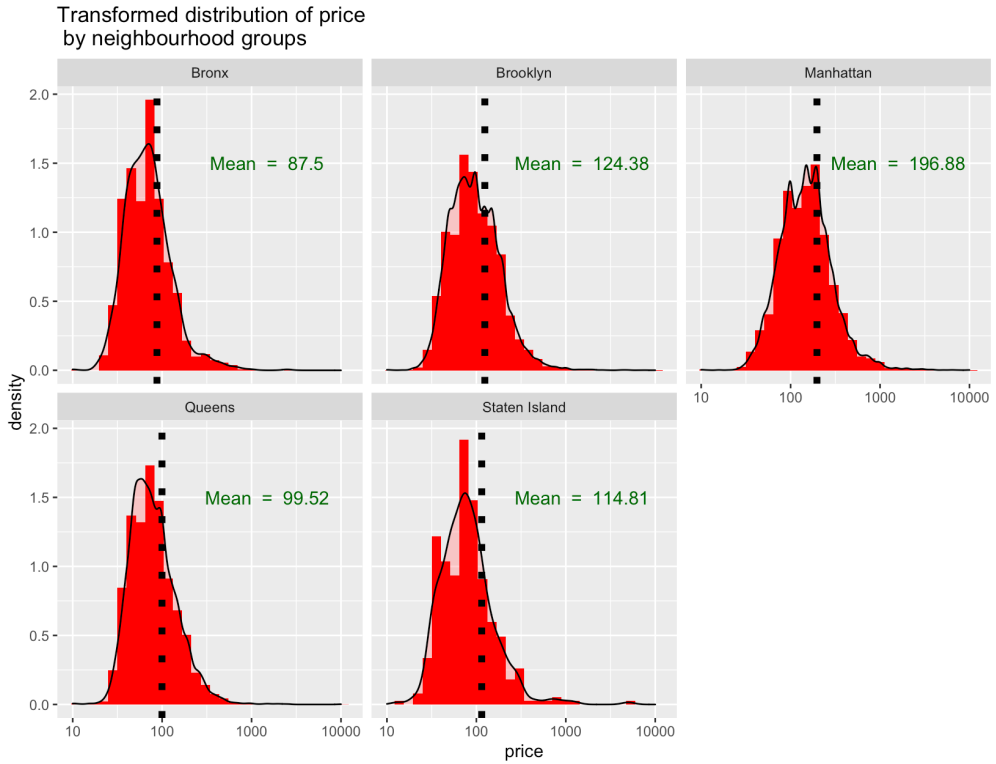
```
airbnb_nh <- airbnb %>%
  group_by(neighbourhood_group) %>%
  summarise(price = round(mean(price), 2))
```

	neighbourhood_group	price
	<fct>	<dbl>
1	Bronx	87.5
2	Brooklyn	124.
3	Manhattan	197.
4	Queens	99.5
5	Staten Island	115.

3. 변수 특징 살펴보기 (시각화)

- neighbourhood_group별 히스토그램 밀도 확인
  1. Manhattan
  2. Brooklyn
  3. Staten Island
  4. Queens
  5. The Bronx

  순서로 가격이 높음




```
ggplot(airbnb, aes(price)) +
  geom_histogram(bins = 30, aes(y = ..density..), fill = "red") +
  geom_density(alpha = 0.2, fill = "red") +
  ggtitle("Transformed distribution of price\n by neighbourhood groups") +
  geom_vline(data = airbnb_nh, aes(xintercept = price), size = 2, linetype = 3) +
  geom_text(data = airbnb_nh,y = 1.5, aes(x = price + 1400, label = paste("Mean  = ",price)), color = "darkgreen", size = 4) +
  facet_wrap(~neighbourhood_group) +
  scale_x_log10()
```

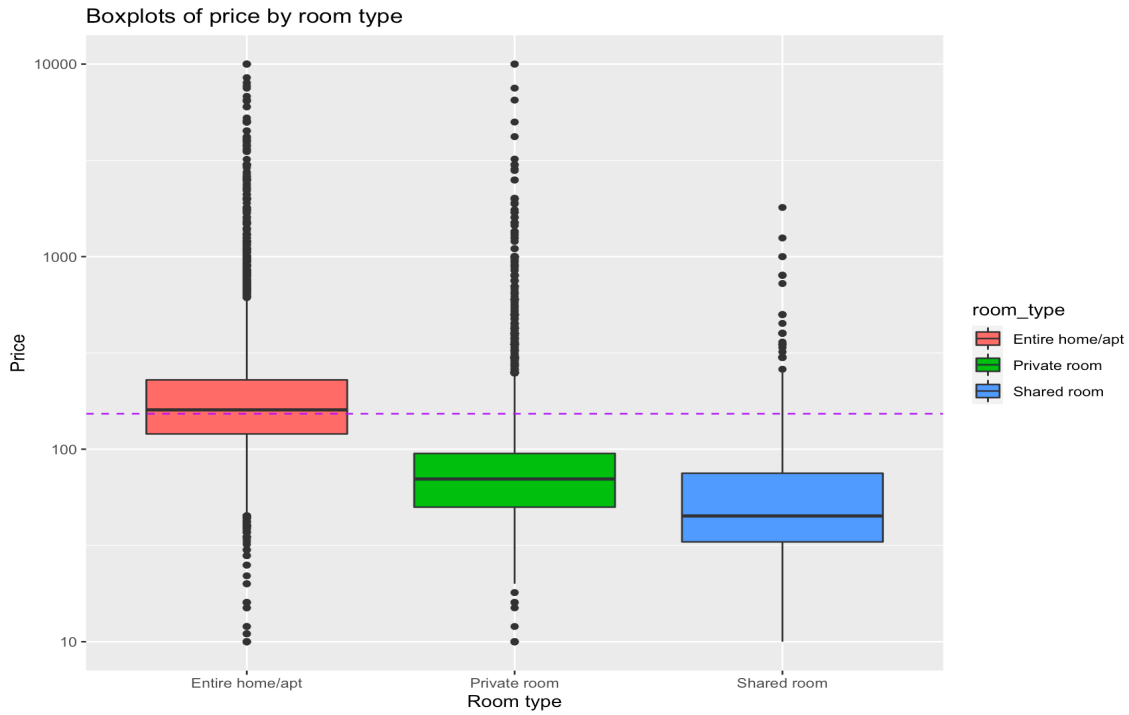
3. 변수 특징 살펴보기 (시각화)

- Room_type별 가격 Boxplot
  1. Entire home or apartment
  2. Private Room
  3. Shared Room

  순서로 가격이 높음

```
> unique(airbnb$room_type)
[1] Private room    Entire home/apt Shared room
Levels: Entire home/apt Private room Shared room
```

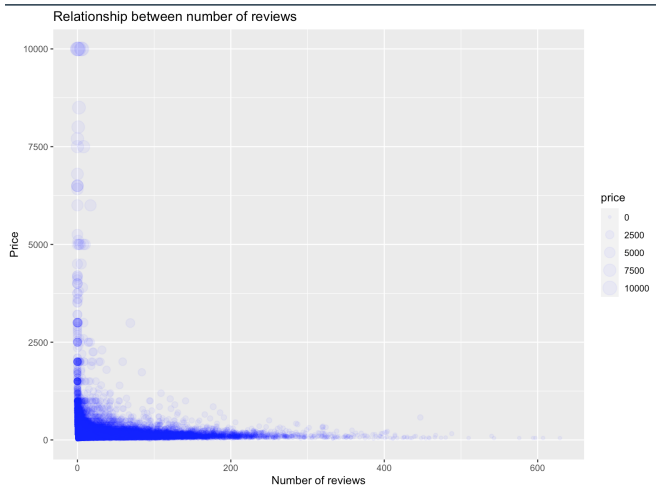
```
ggplot(airbnb, aes(x = room_type, y = price)) +
  geom_boxplot(aes(fill = room_type)) + scale_y_log10()+
  xlab("Room type") +
  ylab("Price") +
  ggtitle("Boxplots of price by room type") +
  geom_hline(yintercept = mean(airbnb$price), color = "purple", linetype = 2)
```



3. 변수 특징 살펴보기 (시각화)

[review 수와 가격 관계]

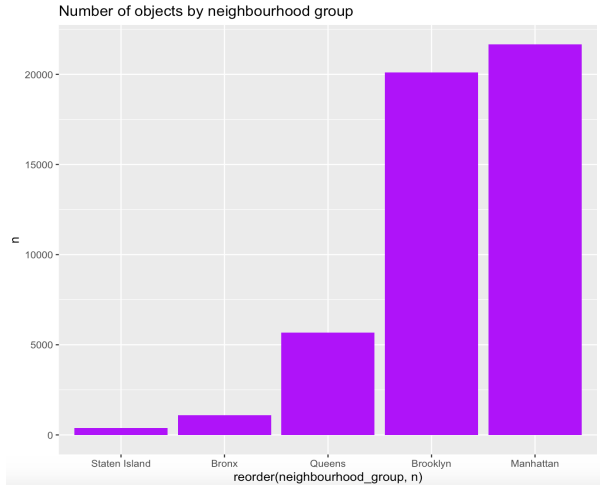
- 가격이 낮을수록 리뷰수가 많았으며 가격이 높을수록 리뷰수가 적음




```
ggplot(airbnb, aes(number_of_reviews, price)) +
  theme(axis.title = element_text(), axis.title.x = element_text()) +
  geom_point(aes(size = price), alpha = 0.05, color = "blue") +
  xlab("Number of reviews") +
  ylab("Price") +
  ggtitle("Relationship between number of reviews",)
```

[neighborhood 그룹별 count]

- 맨하튼에서 airbnb 수가 가장 많이 나타나며 Staten Island에서 airbnb 수가 가장 적음
  - Airbnb 수는 가격에 비례하는 것으로 보임. (예외. Staten Island)

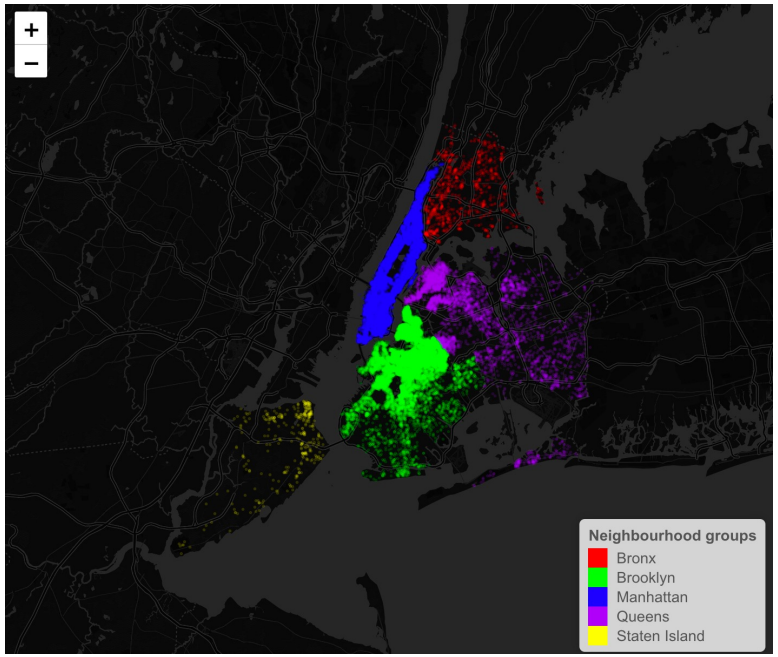


Tally(): summarise()와 같은 기능 → 요약된 열 이름이 n
Reorder: 내림, 오름 차순 변경

```
airbnb %>% group_by(neighbourhood_group) %>% tally() %>%
  ggplot(aes(x = reorder(neighbourhood_group, n), n)) +
  geom_bar(stat = "identity", fill = "purple") +
  ggtitle("Number of objects by neighbourhood group")
```

3. 변수 특징 살펴보기 (시각화)

- 위도 경도 데이터를 이용한 neighbourhood_group별 Airbnb 지도에 표시
  - 지도 시각화
    - install.packages("leaflet")
    - library(leaflet)
  - 맨해튼은 다른 이웃에 비해 좁은 면적을 가지고 있지만 밀도가 높은 것을 보아 Airbnb 수가 다른 지역에 비해 많음
  - 맨해튼, 브루클린은 가격이 비싼 동시에 Airbnb수가 많은 것을 확인할 수 있는 반면 스태튼 아일랜드의 경우 세번째로 가격이 비싼 반면 가장 airbnb 수가 적음



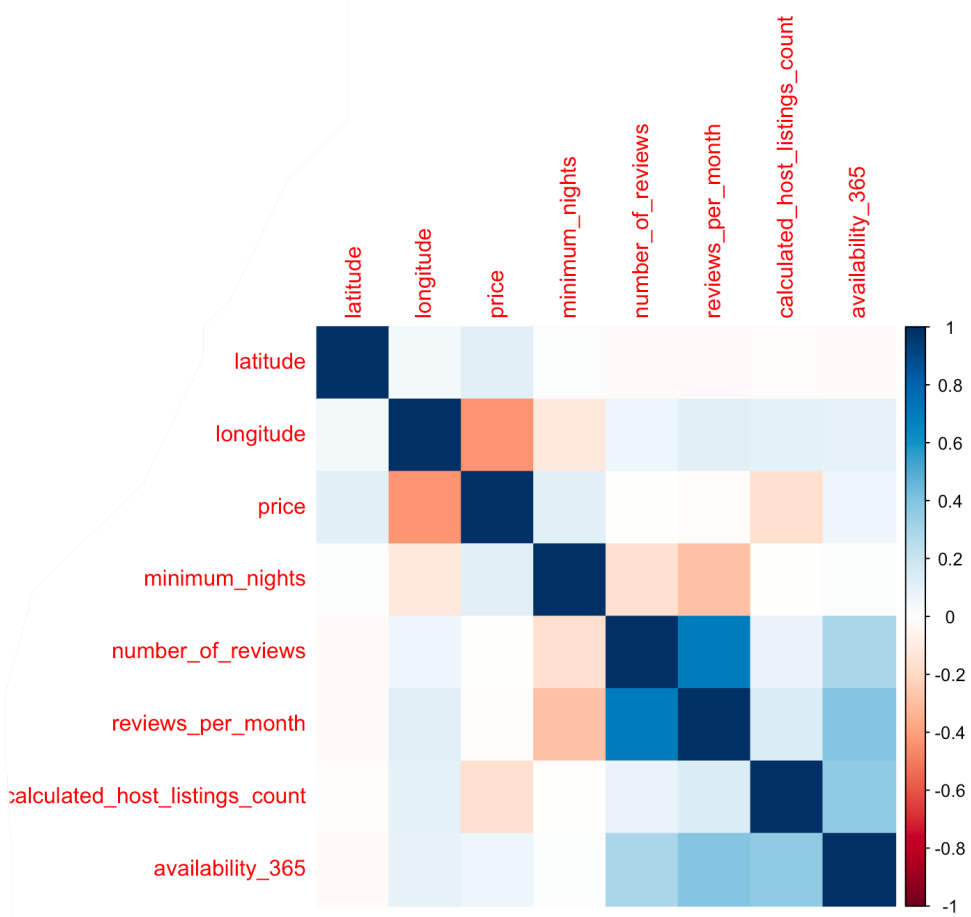

```
install.packages("leaflet")
library(leaflet)

pal <- colorFactor(palette = c("red", "green", "blue", "purple", "yellow"), domain = airbnb$neighbourhood_group)

leaflet(data = airbnb) %>% addProviderTiles(providers$CartoDB.DarkMatterNoLabels) %>%
  addCircleMarkers(~longitude, ~latitude, color = ~pal(neighbourhood_group), weight = 1, radius=1, fillOpacity = 0.1, opacity = 0.1
,label = paste("Name:", airbnb$name))%>% addLegend("bottomright", pal = pal, values = ~neighbourhood_group,title = "Neighbourhood
groups",opacity = 1)
```

addProviderTiles: 지도 불러오기 (provider 이름으로 불러올 수 있음)
addCircleMarkers: 지도 위에 점찍기

4. 변수 간 상관관계



Sapply: 원하는 벡터, 리스트, 데이터프레임에 원하는 함수 적용

```
library(corrplot)
airbnb_cor <- airbnb[, sapply(airbnb, is.numeric)]
airbnb_cor <- airbnb_cor[complete.cases(airbnb_cor), ]
correlation_matrix <- cor(airbnb_cor, method = "spearman")
corrplot(correlation_matrix, method = "color")
```

- Price와 longitude 변수 간의 상관관계가 높아 보임

감사합니다 ☺