XJ

```
## Warning: package 'visdat' was built under R version 3.6.2
```

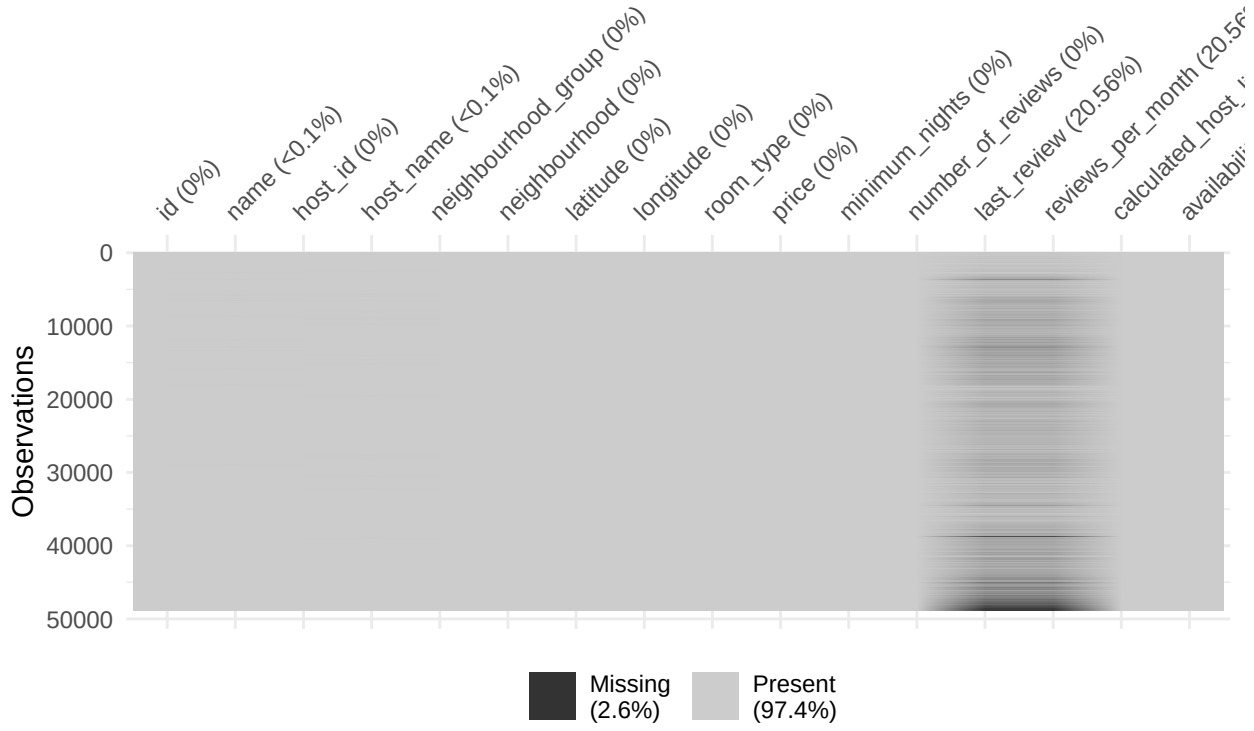
```
## Warning: package 'naniar' was built under R version 3.6.2
```

```
## Warning: package 'corrplot' was built under R version 3.6.2
```

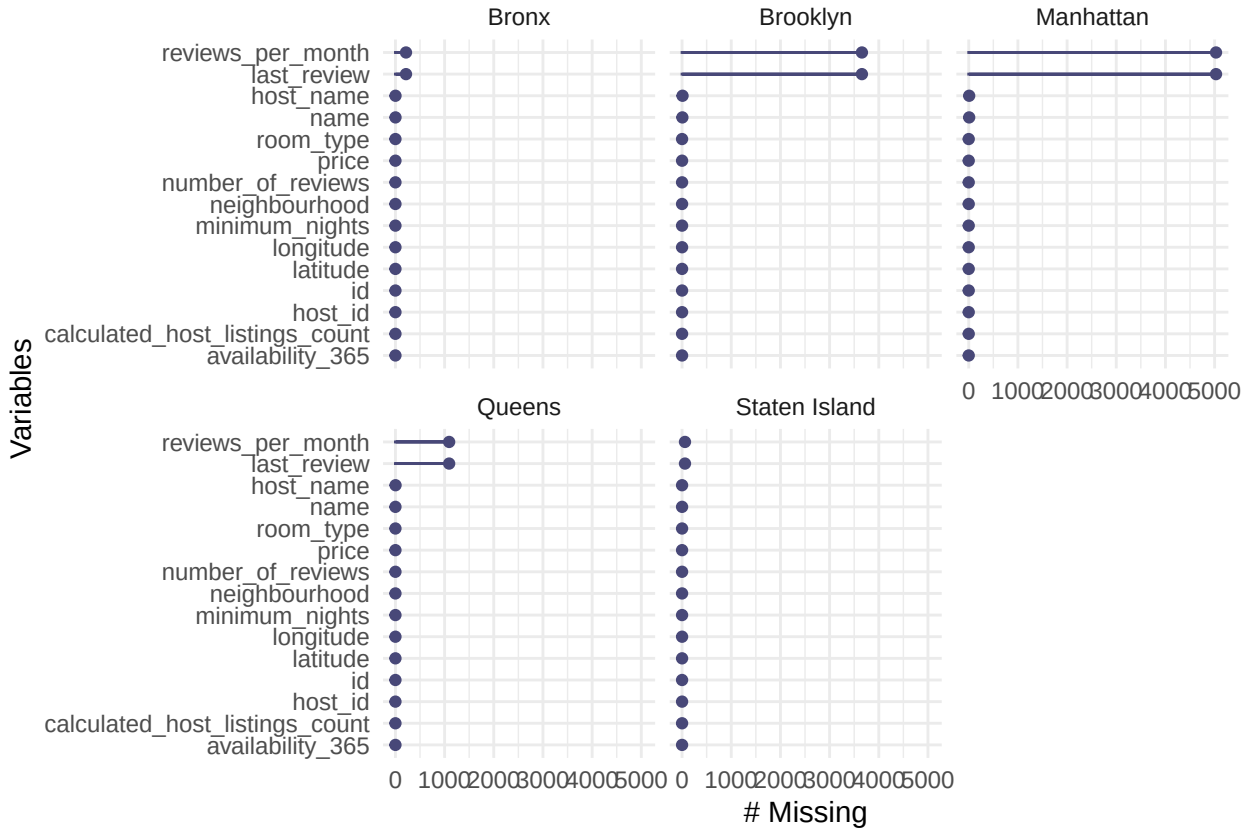
Missing data

- We have about 20% missing in last-review (date) and reviews/ month
  - mainly in Brooklyn and Manhattan
  - impute the data or take complete cases (PMM)
- Modified last_review, as the number of days till 2020-01-01

```
vis_miss(AB_NYC_2019)
```




```
gg_miss_var(AB_NYC_2019,facet = neighbourhood_group) ### evenly missing
```




```
table(AB_NYC_2019$neighbourhood_group)
```

```
##
##         Bronx      Brooklyn     Manhattan        Queens Staten Island
##          1091         20104         21661          5666           373
```

```
AB_NYC_2019$neighbourhood_group<- as.factor(AB_NYC_2019$neighbourhood_group)
AB_NYC_2019$neighbourhood<- as.factor(AB_NYC_2019$neighbourhood)
AB_NYC_2019$room_type<- as.factor(AB_NYC_2019$room_type)

now<- as.Date("2020-01-01")
AB_NYC_2019$last_review<- as.numeric(gsub("([0-9]+).*$", "\\1", now-AB_NYC_2019$last_review))
```

Correlation

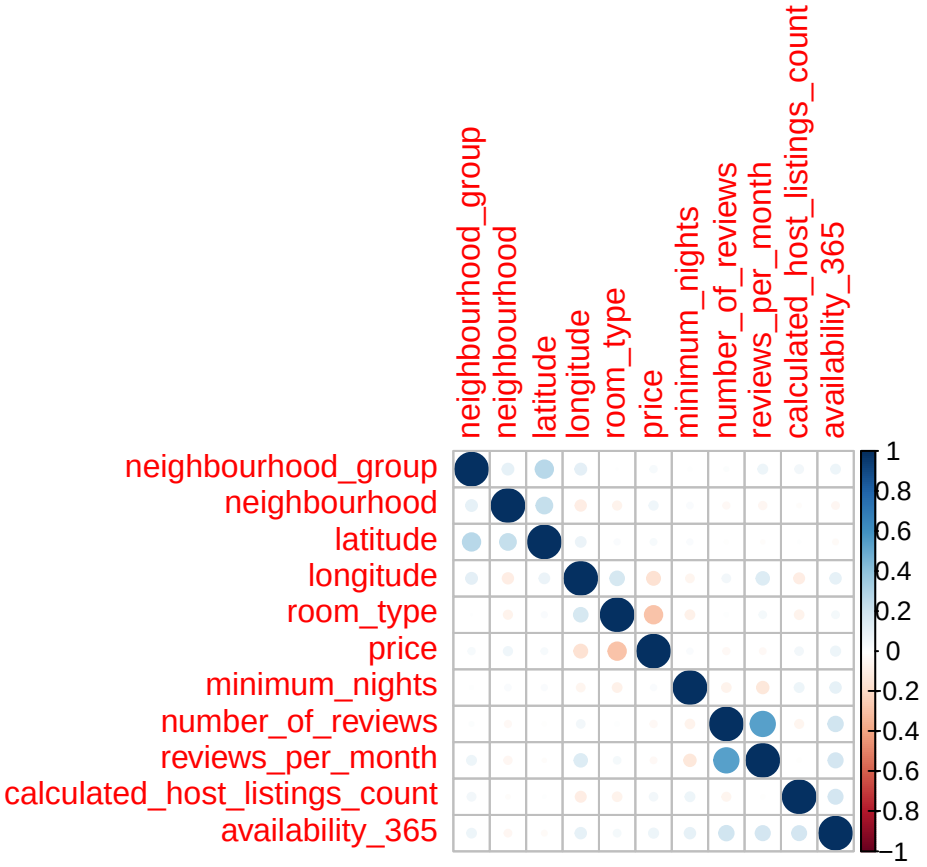
```
corr_check<- AB_NYC_2019 %>% mutate(neighbourhood_group = as.numeric(neighbourhood_group), neighbourhoo
corr_check<- corr_check %>% select(-c(id, name, host_id, host_name, last_review))

corr=cor(as.matrix(na.omit(corr_check)))
corrplot(corr,number.cex=.8, upper="ellipse")
```

```
## Warning in text.default(pos.xlabel[, 1], pos.xlabel[, 2], newcolnames, srt
## = tl.srt, : "upper" is not a graphical parameter
```

```
## Warning in text.default(pos.ylabel[, 1], pos.ylabel[, 2], newrownames, col
## = tl.col, : "upper" is not a graphical parameter
```

```
## Warning in title(title, ...): "upper" is not a graphical parameter
```



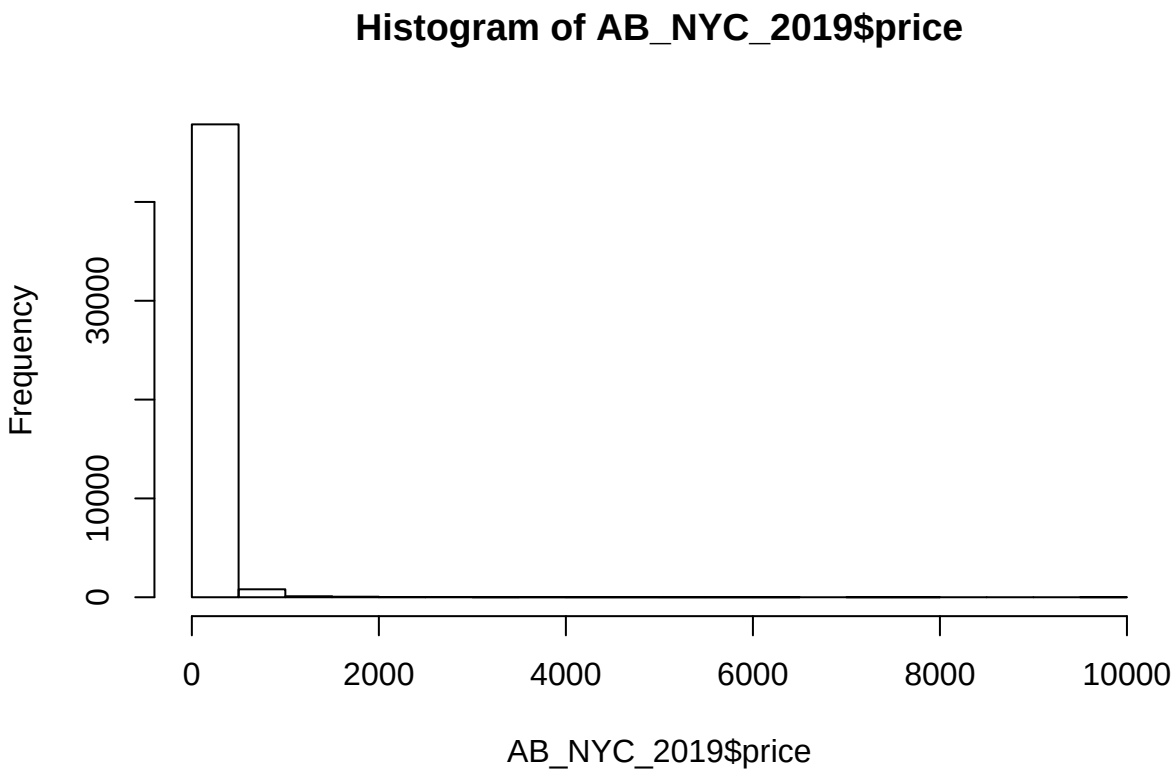
plots for covariates

- Distribution for price is right skewed, could us log(price)
- could use reviews/ month to quantify popularity
  - Again, right skewed, could use log scale

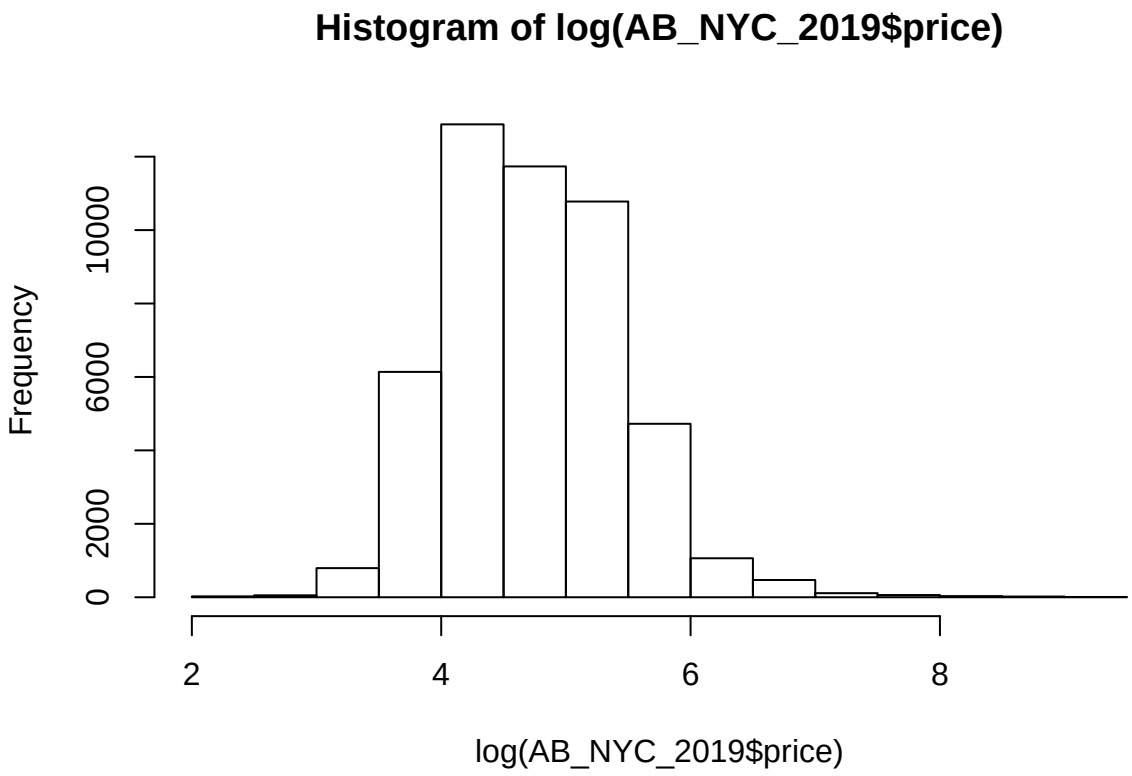
```
max(AB_NYC_2019$price)
```

```
## [1] 10000
```

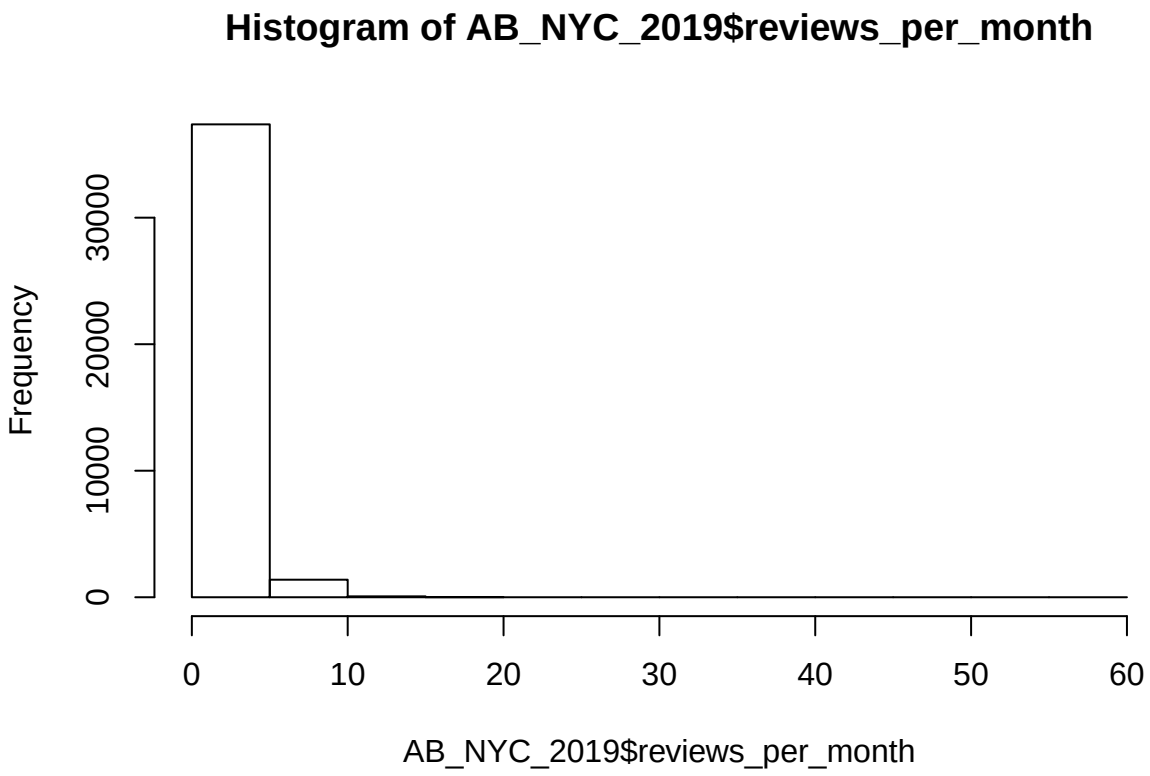
```
hist(AB_NYC_2019$price)
```



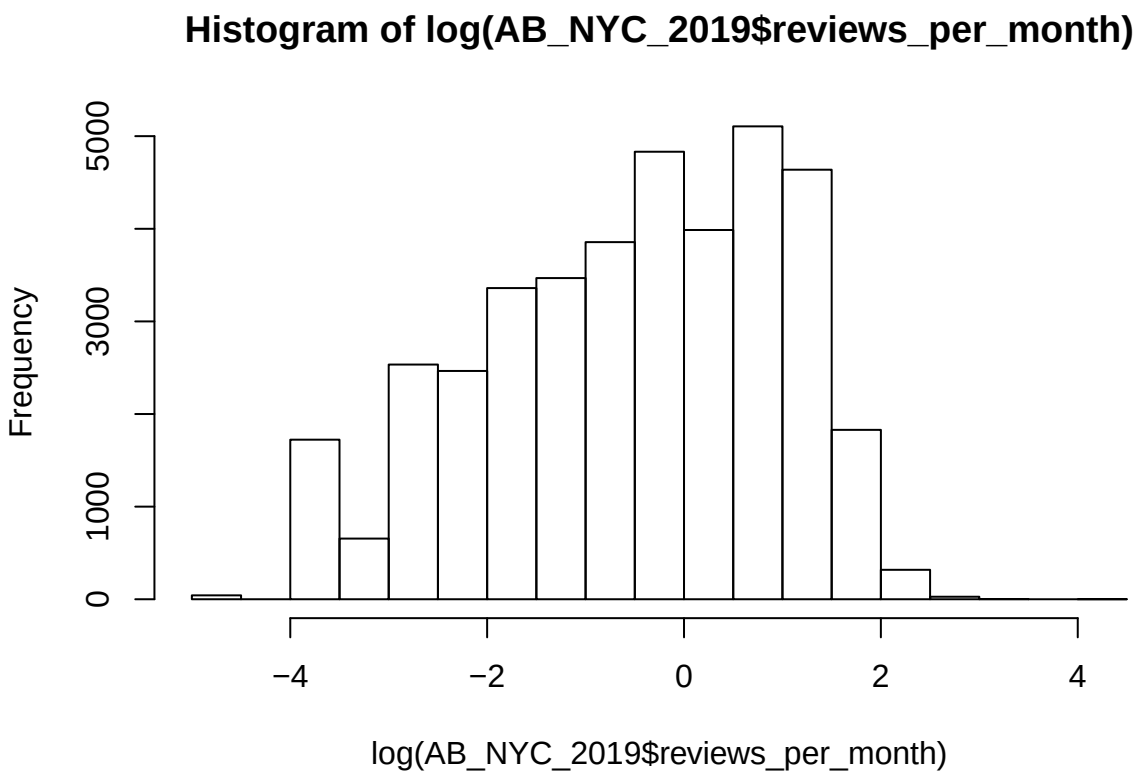

```
hist(log(AB_NYC_2019$price))
```



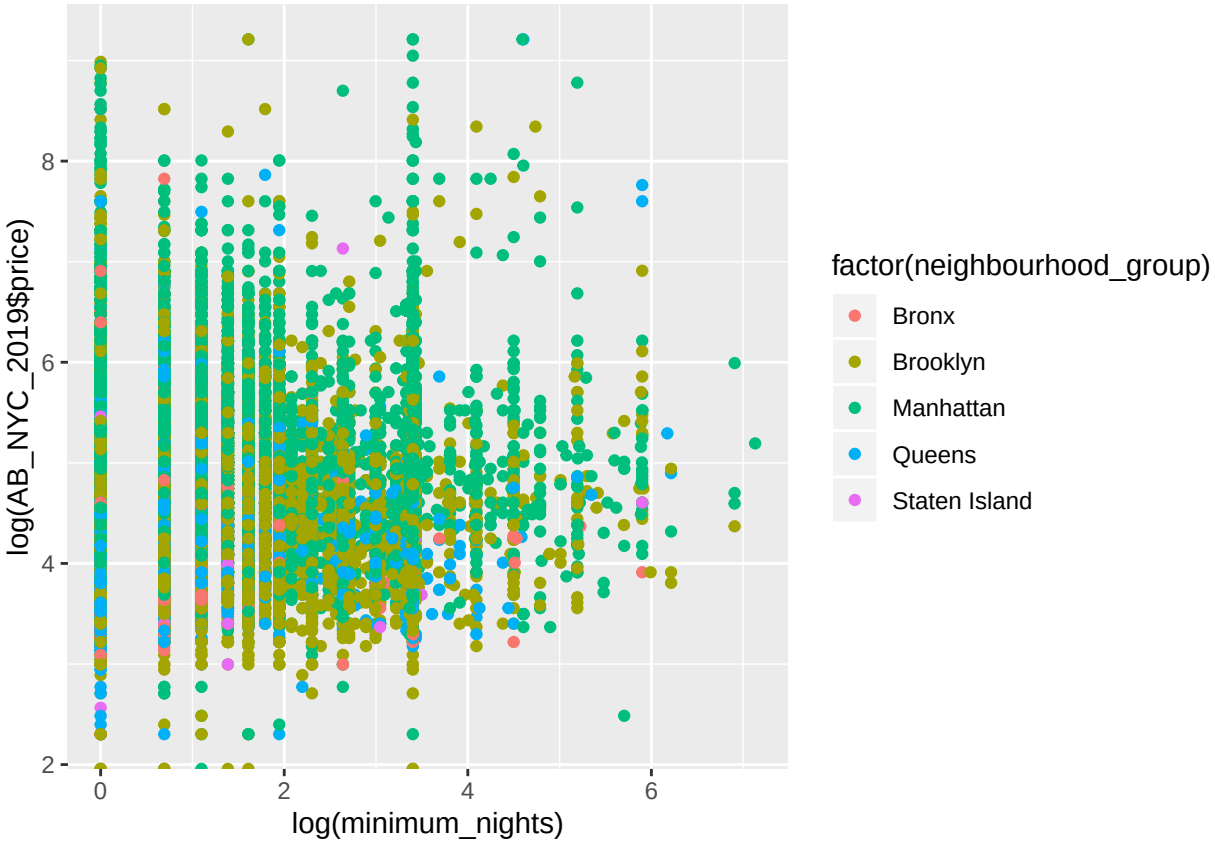

```r
hist(AB_NYC_2019$reviews_per_month)
```



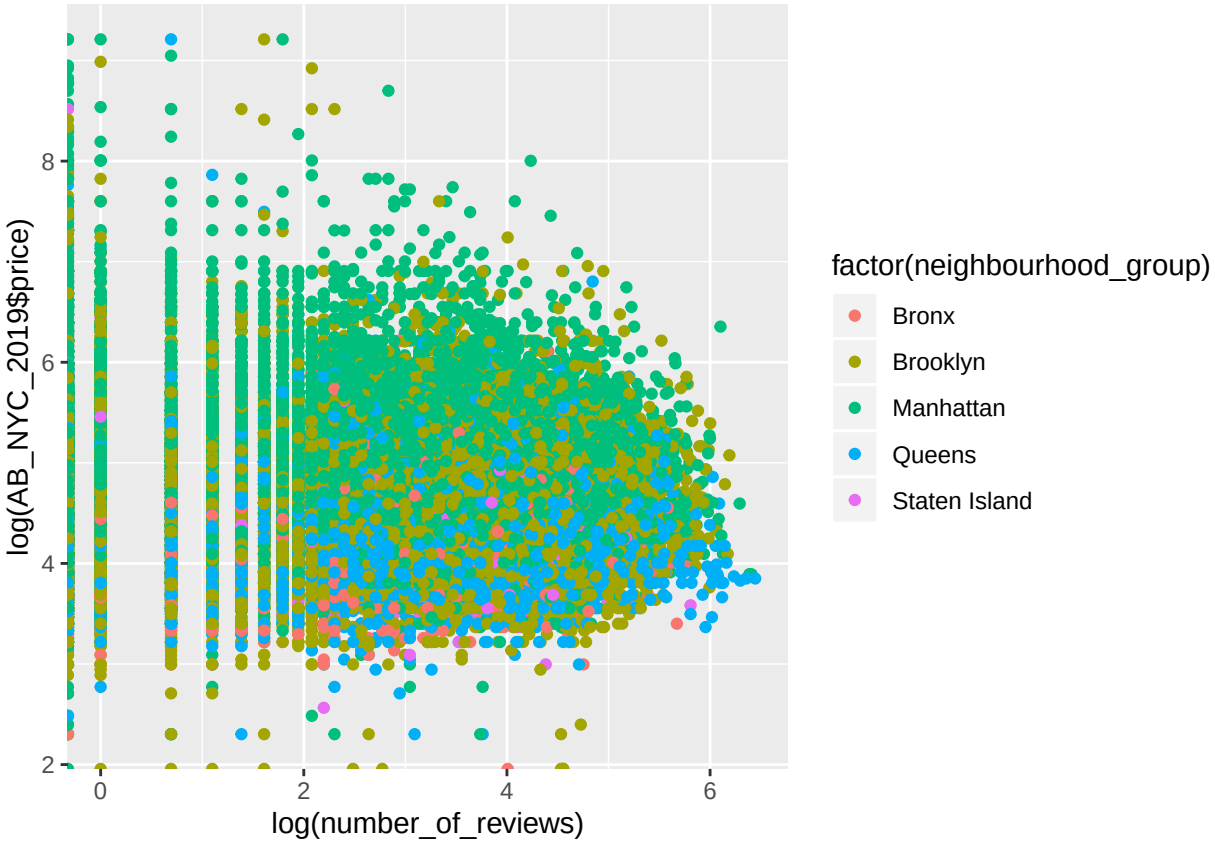

```
hist(log(AB_NYC_2019$reviews_per_month))
```



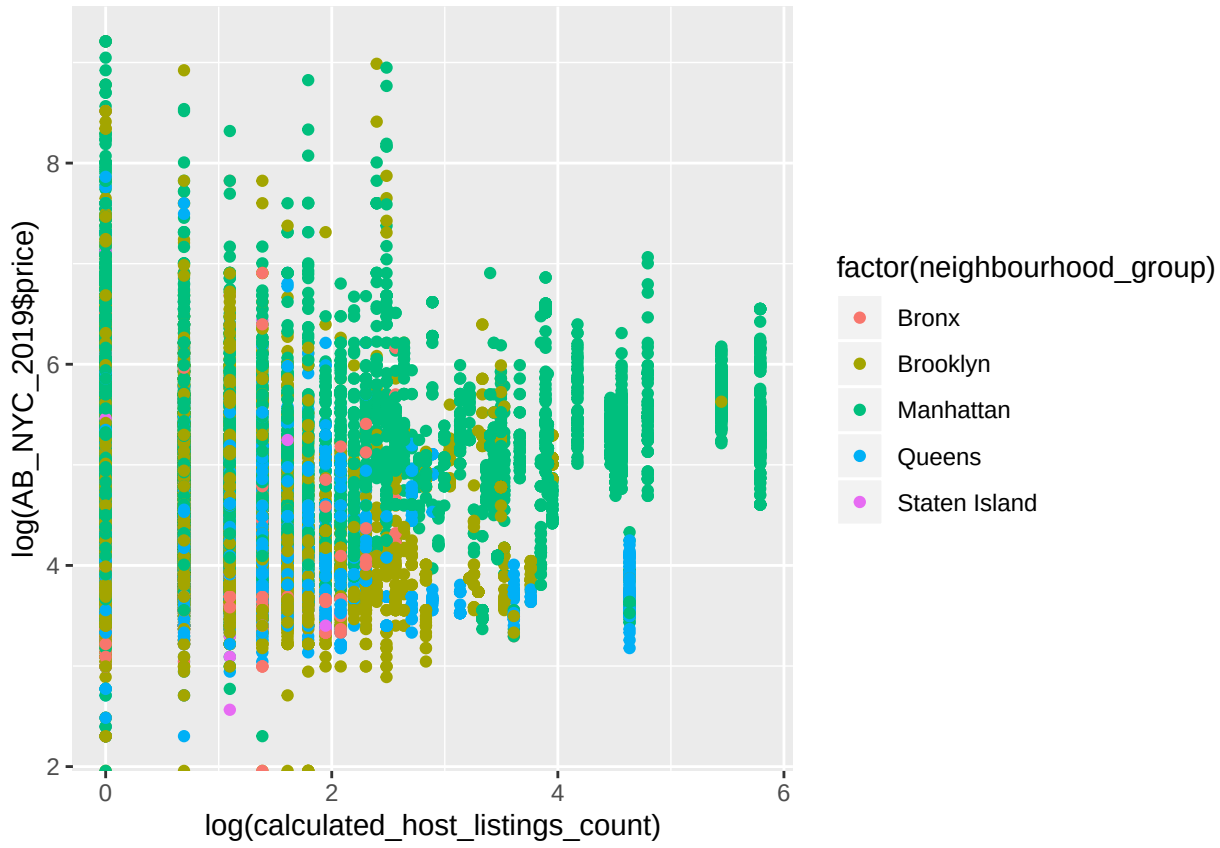

```
### Price as response
qplot(x = log(minimum_nights), y = log(AB_NYC_2019$price), data=AB_NYC_2019, color = factor(neighbourho
```




```
qplot(x = log(number_of_reviews), y = log(AB_NYC_2019$price), data=AB_NYC_2019, color = factor(neighbou
```




```
qplot(x = log(calculated_host_listings_count), y = log(AB_NYC_2019$price), data=AB_NYC_2019, color = fa
```

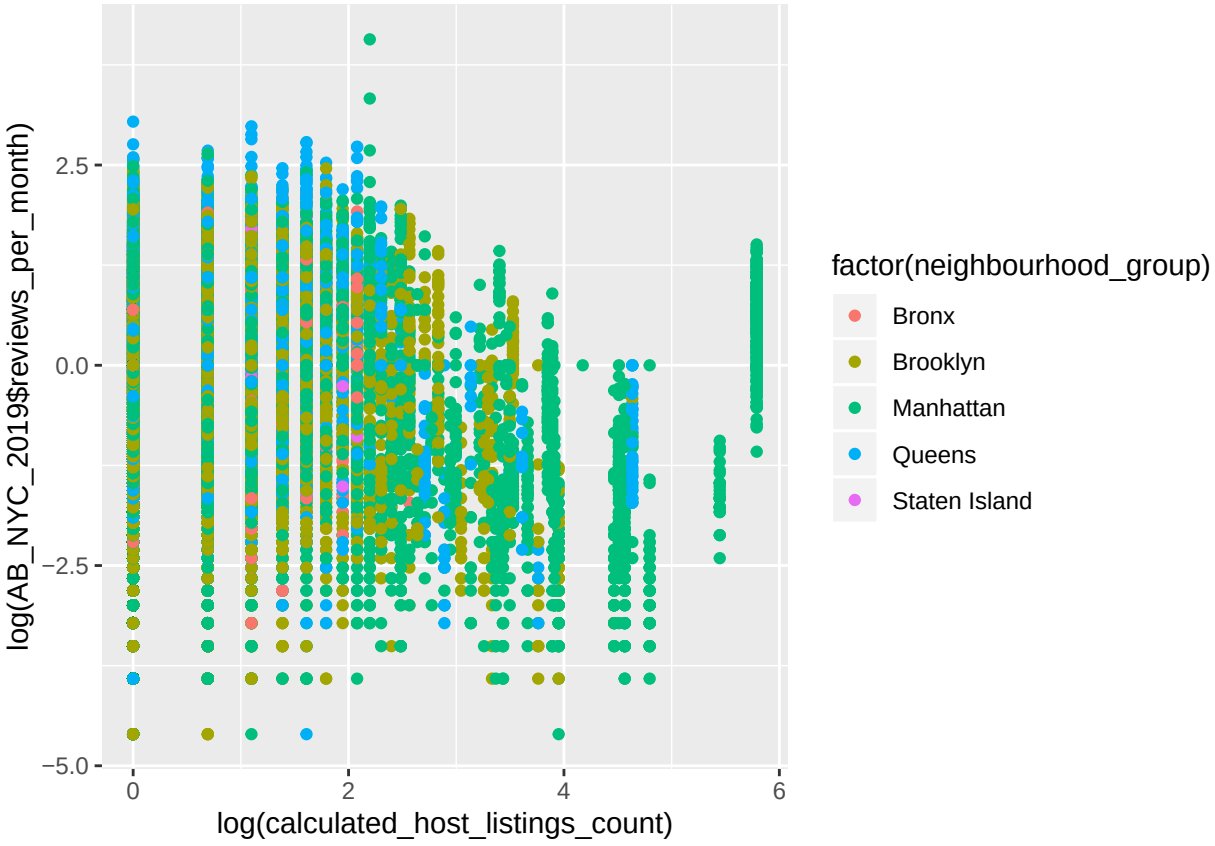



```
### reviews/ month as response
qplot(x = log(minimum_nights), y = log(AB_NYC_2019$reviews_per_month), data=AB_NYC_2019, color = factor
```

```
## Warning: Removed 10052 rows containing missing values (geom_point).
```

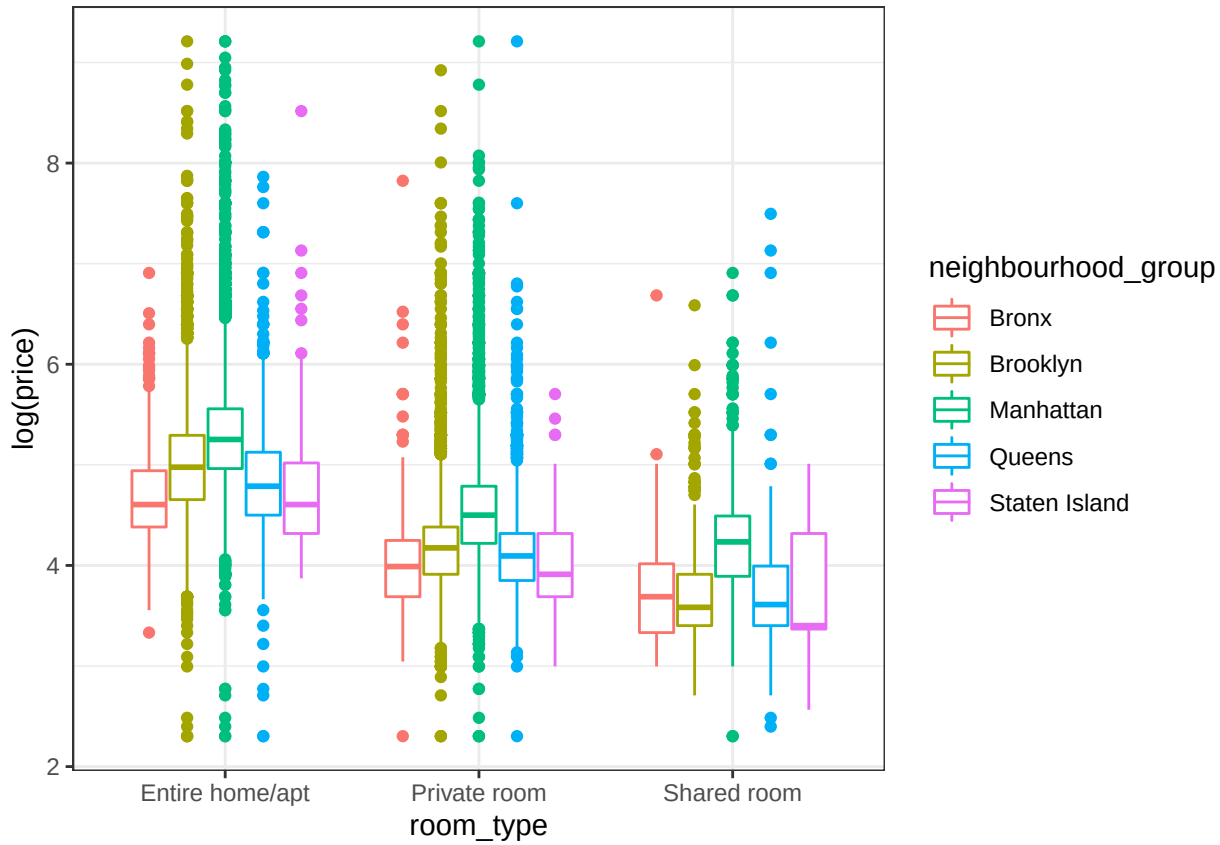

```
qplot(x = log(calculated_host_listings_count), y = log(AB_NYC_2019$reviews_per_month), data=AB_NYC_2019
```

```
## Warning: Removed 10052 rows containing missing values (geom_point).
```



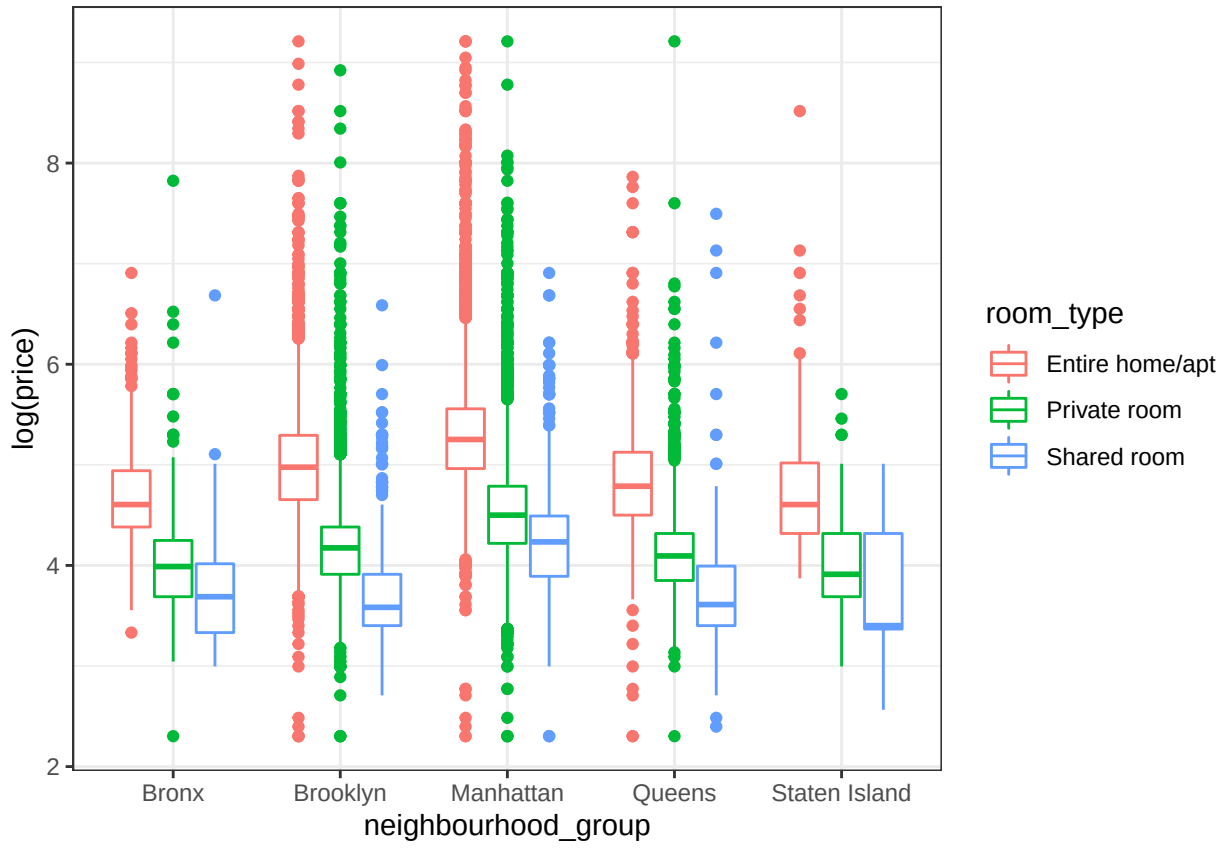

```
ggplot(AB_NYC_2019,aes(x=room_type,y=log(price),color=neighbourhood_group))+geom_boxplot()+theme_bw()
```

```
## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
```



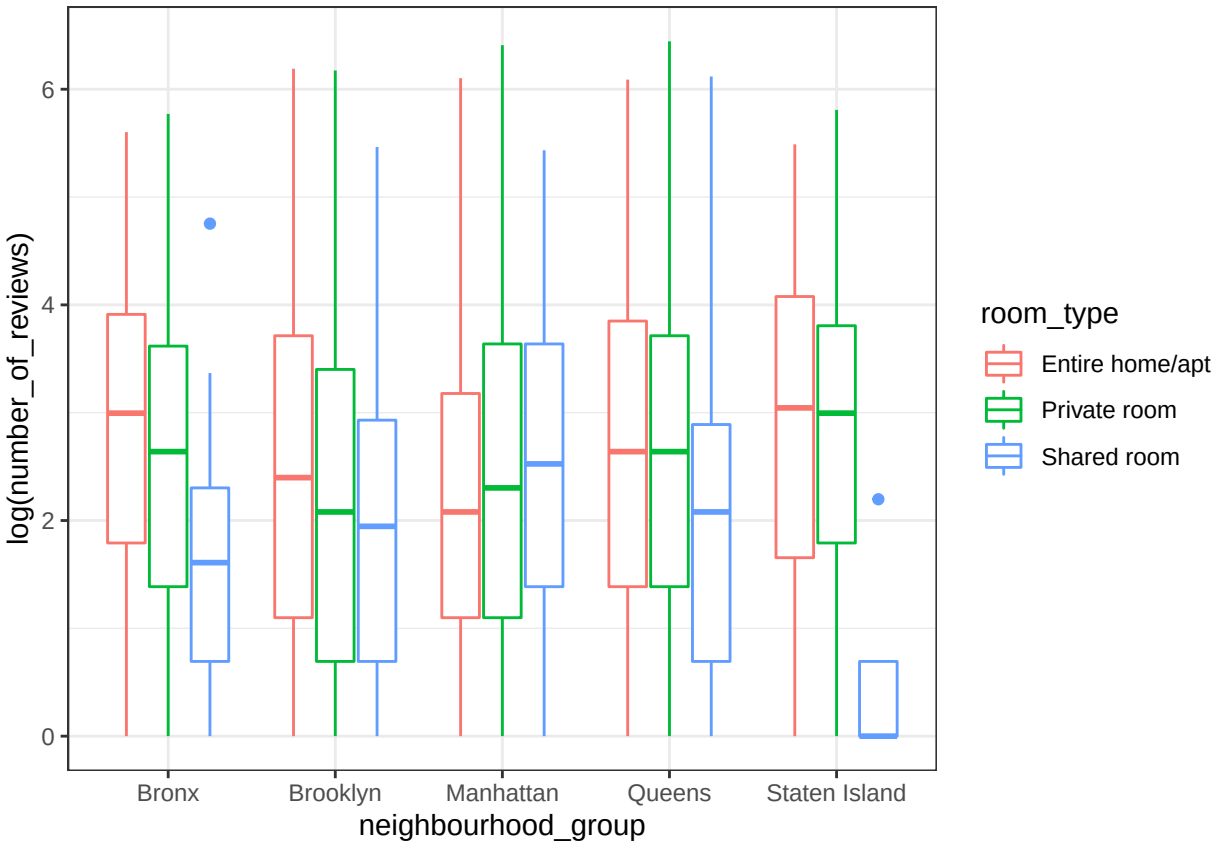

```
ggplot(AB_NYC_2019,aes(x=neighbourhood_group,y=log(price),color=room_type))+geom_boxplot()+theme_bw()
```

```
## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
```




```
ggplot(AB_NYC_2019,aes(x=neighbourhood_group,y=log(number_of_reviews),color=room_type))+geom_boxplot()+
```

```
## Warning: Removed 10052 rows containing non-finite values (stat_boxplot).
```


log(number_of_reviews)
6
4
2
0
Bronx
Brooklyn
Manhattan
Queens
Staten Island
neighbourhood_group
room_type
Entire home/apt
Private room
Shared room