

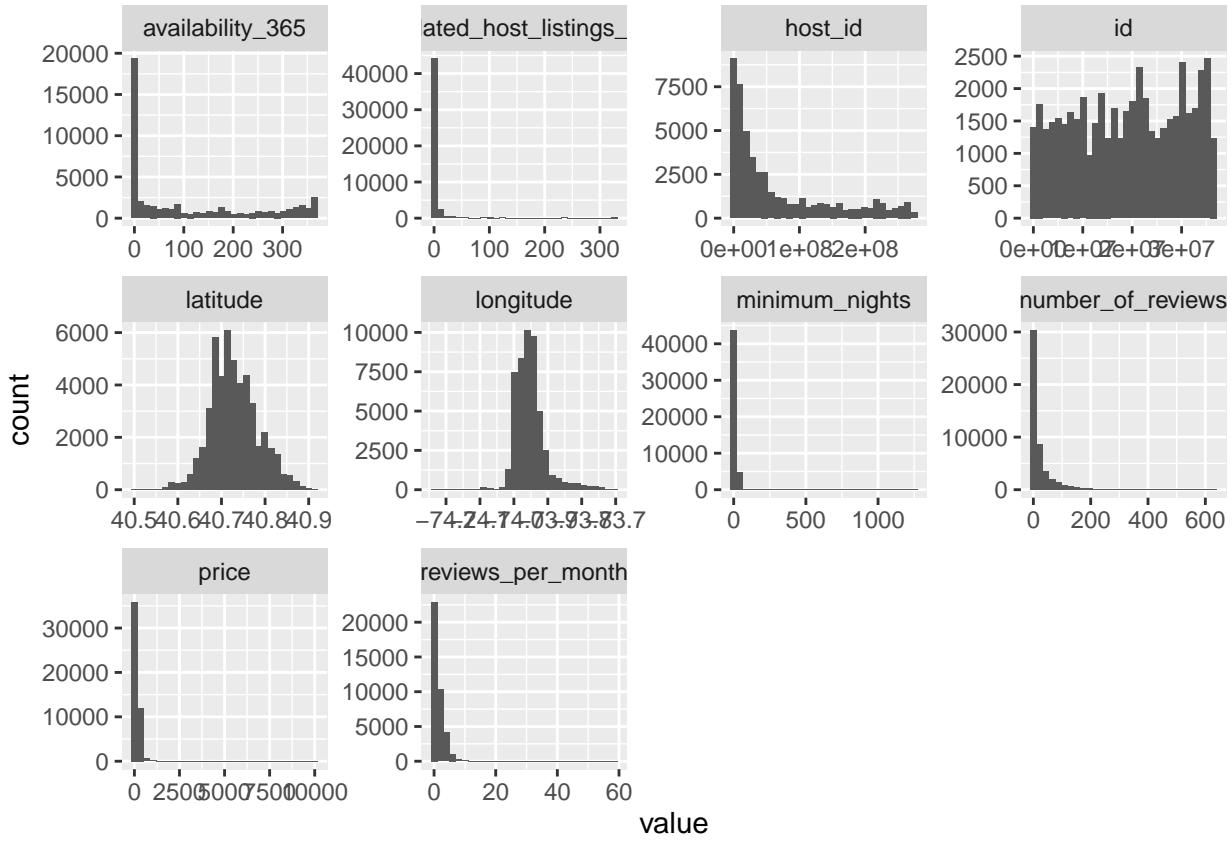
EDA

```
dat = read.csv('AB_NYC_2019.csv', na.strings = c("", "NA"))

library(purrr)
library(tidyr)
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.5.2
dat %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~ key, scales = "free") +
  geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 10052 rows containing non-finite values (stat_bin).
```



```
lapply(dat, class)
```

```
## $id
## [1] "integer"
##
## $name
## [1] "factor"
```

```

## 
## $host_id
## [1] "integer"
##
## $host_name
## [1] "factor"
##
## $neighbourhood_group
## [1] "factor"
##
## $neighbourhood
## [1] "factor"
##
## $latitude
## [1] "numeric"
##
## $longitude
## [1] "numeric"
##
## $room_type
## [1] "factor"
##
## $price
## [1] "integer"
##
## $minimum_nights
## [1] "integer"
##
## $number_of_reviews
## [1] "integer"
##
## $last_review
## [1] "factor"
##
## $reviews_per_month
## [1] "numeric"
##
## $calculated_host_listings_count
## [1] "integer"
##
## $availability_365
## [1] "integer"

```

Missing Data Manipulation

```
apply(dat, 2, function(x) (sum(is.na(x))))
```

##	id	name
##	0	16
##	host_id	host_name
##	0	21
##	neighbourhood_group	neighbourhood
##	0	0

```

##           latitude          longitude
##                  0                   0
##           room_type          price
##                  0                   0
##      minimum_nights number_of_reviews
##                  0                   0
##           last_review reviews_per_month
##                  10052                10052
## calculated_host_listings_count availability_365
##                               0                   0

dat = dat[, !names(dat) %in% c('id', 'host_name', 'last_review')]
dat$reviews_per_month[is.na(dat$reviews_per_month)] = 0
apply(dat, 2, function(x)(sum(is.na(x))))


##           name          host_id
##                  16                   0
## neighbourhood_group neighbourhood
##                  0                   0
##           latitude          longitude
##                  0                   0
##           room_type          price
##                  0                   0
##      minimum_nights number_of_reviews
##                  0                   0
## reviews_per_month calculated_host_listings_count
##                  0                   0
##      availability_365
##                               0
##
```

EDA plots

```

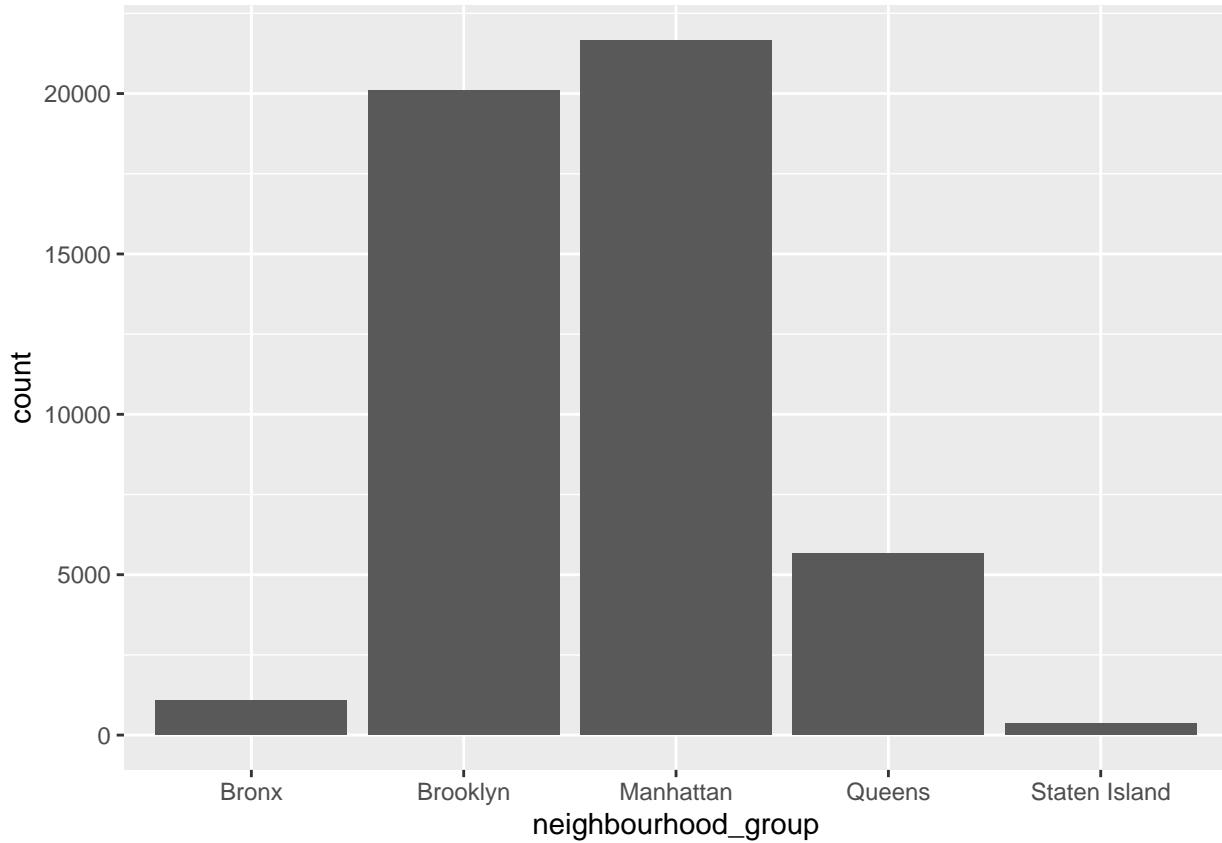
library(jpeg)
library(ggpubr)

## Warning: package 'ggpubr' was built under R version 3.5.2
## Loading required package: magrittr
##
## Attaching package: 'magrittr'
## The following object is masked from 'package:tidyverse':
##     extract
## The following object is masked from 'package:purrr':
##     set_names
library(grid)

img = readJPEG("New_York_City_.jpg")

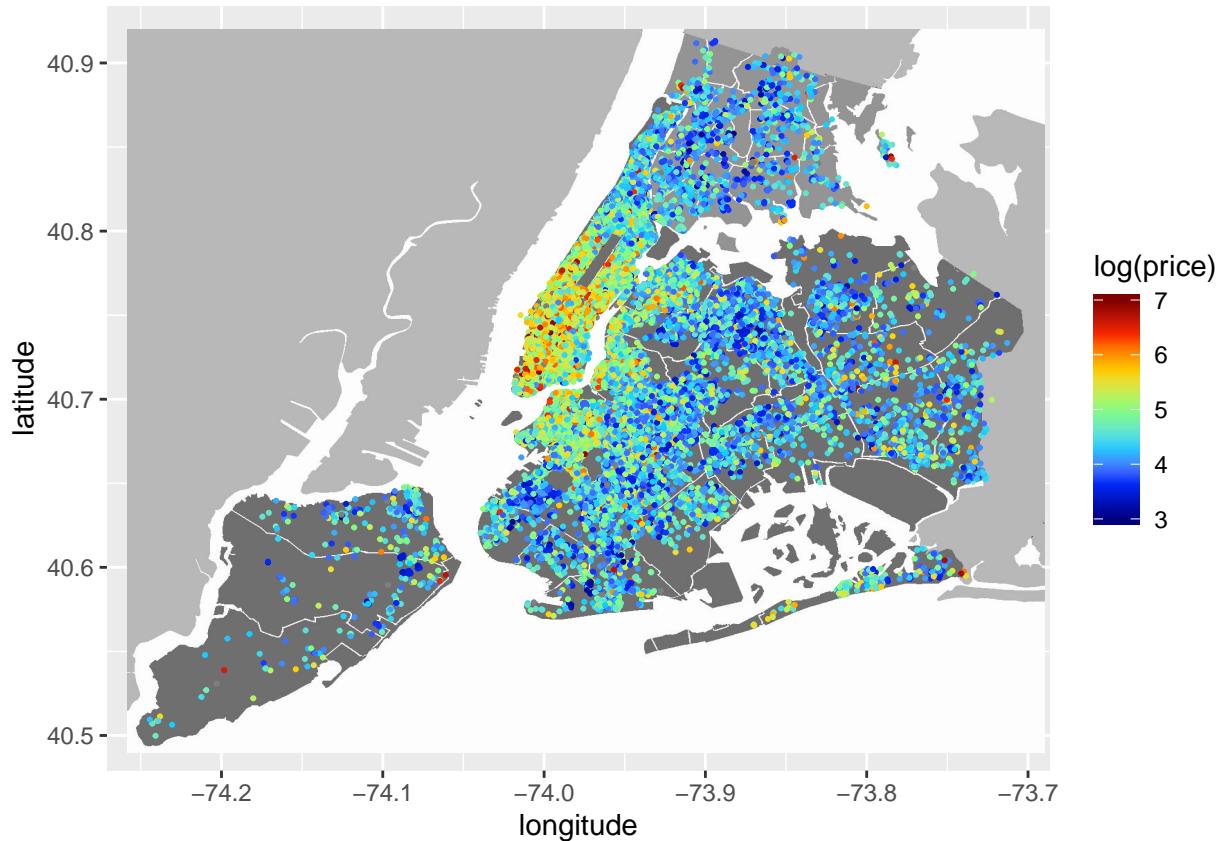
```

```
ggplot(dat, aes(x=neighbourhood_group)) +
  geom_bar()
```

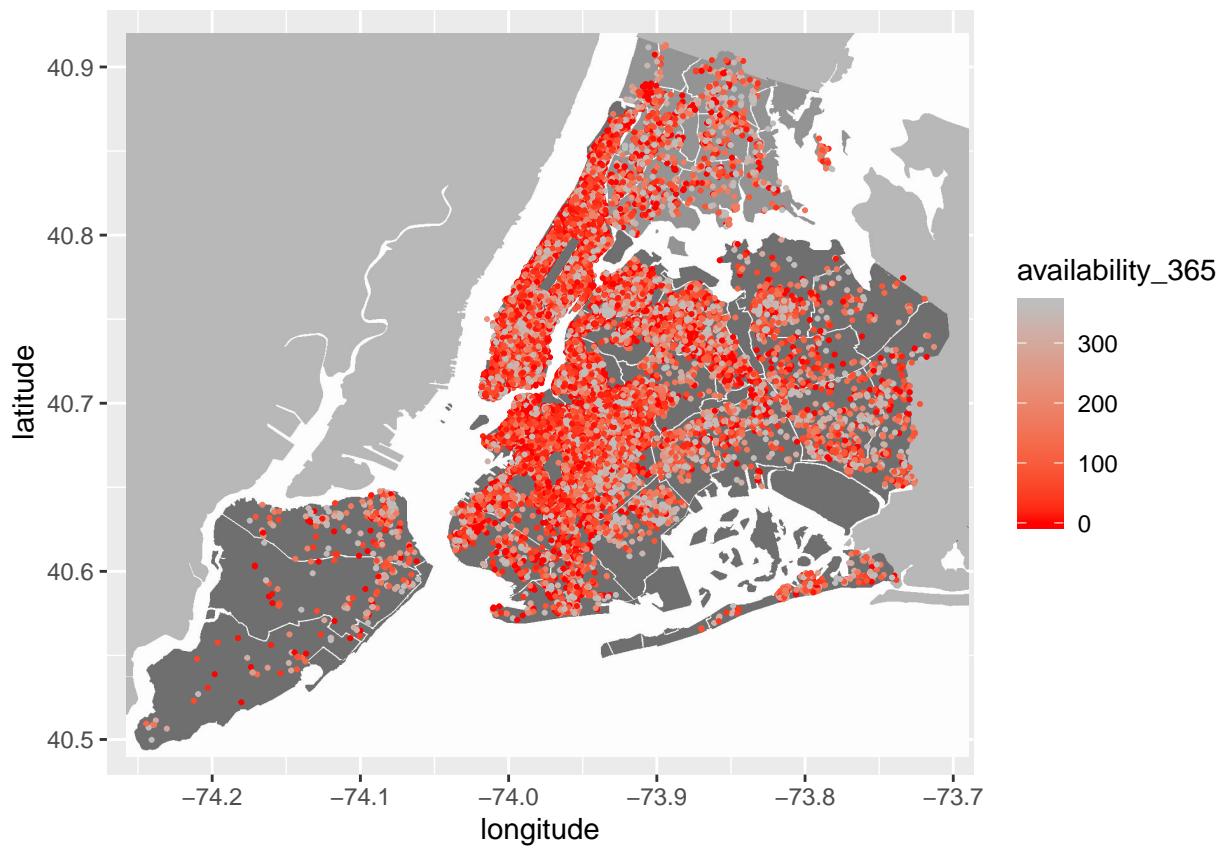


```
jet.colors <- colorRampPalette(c("#00007F", "blue", "#007FFF", "cyan", "#7FFF7F", "yellow", "#FF7F00", "#FF0000"))

ggplot(dat, aes(x=longitude, y = latitude, color = log(price))) +
  annotation_custom(rasterGrob(img,
                                width = unit(1,"npc"),
                                height = unit(1,"npc")),
                     -74.258, -73.69, 40.49,40.92) +
  geom_point(cex = 0.4) +
  scale_colour_gradientn(colors = jet.colors(7), limits = c(3,7))
```



```
ggplot(dat, aes(x=longitude, y = latitude, color = availability_365))+
  annotation_custom(rasterGrob(img,
                                width = unit(1, "npc"),
                                height = unit(1, "npc")),
                    -74.258, -73.69, 40.49,40.92) +
  geom_point(cex = 0.4) +
  scale_colour_gradient(low = 'red', high = 'grey')
```



```
ggplot(dat, aes(x=longitude, y = latitude, color = neighbourhood_group))+  
  annotation_custom(rasterGrob(img,  
                                width = unit(1, "npc"),  
                                height = unit(1, "npc")),  
                    -74.258, -73.69, 40.49,40.92) +  
  geom_point(cex = 0.4)
```

