Descriptive analysis - Airbnb at New York in 2019

Descriptive analysis

Data cleaning

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
df <- read.csv("~/Documents/SPR21/STA160/Airbnb_NYC_2019.csv")
df$last_review <- ymd(df$last_review)
df$id <- str_pad(df$id,8,pad = "0")
df$host_id <- str_pad(df$host_id,9,pad = "0")
is.na(df) <- df == ''
dfRemove <- df %>%
  filter(year(df$last_review) < "2019")
dfNew <- setdiff(df, dfRemove)
dfRemove <- dfNew %>%
  filter(dfNew$availability_365 >= 365)
dfNew <- setdiff(dfNew, dfRemove)
dfRemove <- dfNew %>%
  filter(is.na(last_review) & availability_365 == 1)
dfNew <- setdiff(dfNew, dfRemove)
dfRemove <- dfNew %>%
  filter(price < 35 | price > 500)
dfNew <- setdiff(dfNew, dfRemove)
```

New dataset after removing outliers

Descriptive analysis

```
dfNew$name <- as.factor(dfNew$name)
dfNew$host_id <- as.factor(dfNew$host_id)
dfNew$host_name <- as.factor(dfNew$host_name)
dfNew$neighbourhood_group <- as.factor(dfNew$neighbourhood_group)
dfNew$neighbourhood <- as.factor(dfNew$neighbourhood)
dfNew$room_type <- as.factor(dfNew$room_type)
summary(dfNew)
```

```
##       id                                                  name
##  Length:32809       Home away from home                    :   13
##  Class :character   Artsy Private BR in Fort Greene Cumberland:   10
##  Mode  :character   Loft Suite @ The Box House Hotel       :   10
##                     New york Multi-unit building           :    9
##                     Brooklyn Apartment                     :    8
##                     (Other)                                :32750
##                     NA's                                   :    9
##       host_id             host_name        neighbourhood_group
##  219517861:  303   Sonder (NYC):  303   Bronx        :  796
##  107434423:  223   Michael     :  276   Brooklyn     :13277
##  137358866:   84   David       :  267   Manhattan    :14286
##  012243051:   81   Blueground  :  223   Queens       : 4170
##  061391963:   68   John        :  208   Staten Island:  280
##  030283594:   66   (Other)     :31519
##  (Other)  :31984   NA's        :   13
##             neighbourhood      latitude       longitude
##  Bedford-Stuyvesant: 2615   Min.   :40.50   Min.   :-74.24
##  Williamsburg      : 2529   1st Qu.:40.69   1st Qu.:-73.98
##  Harlem            : 1802   Median :40.72   Median :-73.95
##  Bushwick          : 1618   Mean   :40.73   Mean   :-73.95
##  Hell's Kitchen    : 1439   3rd Qu.:40.76   3rd Qu.:-73.93
##  East Village      : 1180   Max.   :40.91   Max.   :-73.71
##  (Other)           :21626
##           room_type         price       minimum_nights     number_of_reviews
##  Entire home/apt:17277   Min.   : 35.0   Min.   :   1.000   Min.   :  0.00
##  Private room   :14961   1st Qu.: 70.0   1st Qu.:   1.000   1st Qu.:  0.00
##  Shared room    :  571   Median :110.0   Median :   2.000   Median :  8.00
##                          Mean   :136.2   Mean   :   6.398   Mean   : 29.86
##                          3rd Qu.:176.0   3rd Qu.:   4.000   3rd Qu.: 36.00
##                          Max.   :500.0   Max.   :1000.000   Max.   :629.00
##
##   last_review         reviews_per_month calculated_host_listings_count
##  Min.   :2019-01-01   Min.   : 0.020    Min.   :  1.000
##  1st Qu.:2019-05-24   1st Qu.: 0.670    1st Qu.:  1.000
##  Median :2019-06-19   Median : 1.490    Median :  1.000
##  Mean   :2019-05-30   Mean   : 1.996    Mean   :  8.399
##  3rd Qu.:2019-06-29   3rd Qu.: 2.870    3rd Qu.:  2.000
##  Max.   :2019-07-08   Max.   :58.500    Max.   :327.000
##  NA's   :8965         NA's   :8965
```

```
##  availability_365
##  Min.   :  0.0
##  1st Qu.:  3.0
##  Median : 82.0
##  Mean   :127.3
##  3rd Qu.:247.0
##  Max.   :364.0
##
```

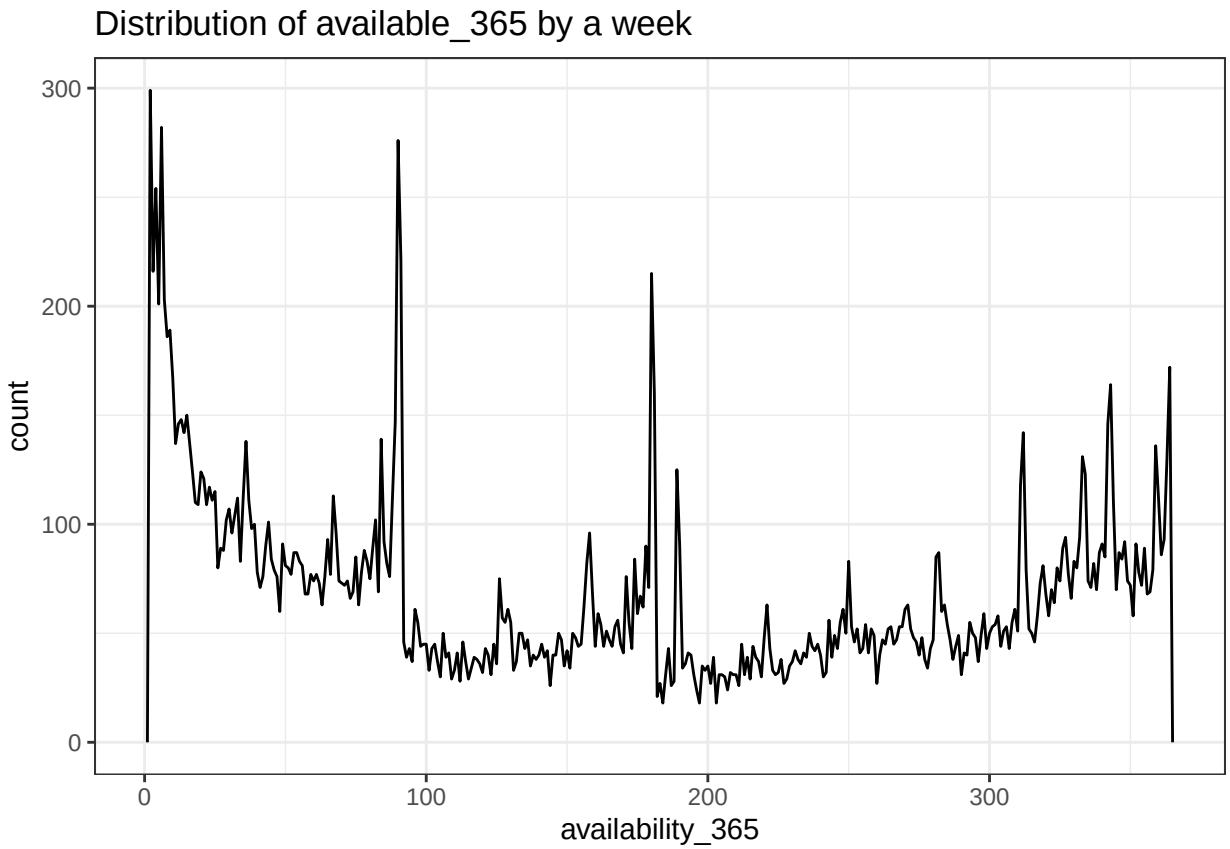
We have 32809 observations, id is the unique key. Name is not focus so we can not choose as catergorical variables Categorical variables: neighbourhood_group, neighbourhood, room_type Host_name: Sonder, Michael, David, Blueground have a chain leasing. ### Zoom in figures #### Available 365

```
dfNew$availability_365 <- as.factor(dfNew$availability_365)
dfNew%>%
  select(id,neighbourhood_group,availability_365) %>%
  summary()
```

```
##       id              neighbourhood_group availability_365
##  Length:32809       Bronx         :  796    0      : 7598
##  Class :character   Brooklyn      :13277    364    :  341
##  Mode  :character   Manhattan     :14286    1      :  299
##                     Queens        : 4170    5      :  282
##                     Staten Island:  280    89     :  276
##                                            3      :  254
##                                            (Other):23759
```

In the top 6 availbability_365, the gap between availability = 1 and the others so different, so I make new dataset with available_365 > 1

```
dfNew$availability_365 <- as.numeric(dfNew$availability_365)
dfNew %>%  filter((availability_365 > 1 & availability_365 < 365) ) %>%
  ggplot(aes(x=availability_365)) +
  geom_freqpoly(binwidth = 1)+
  labs(title = "Distribution of available_365 by a week") +
  theme_bw()
```



As we can see available are high at <5 days, 90 days, 180 days, and 365 days. So there relationshiop between available and minimunights, because the numbers are the same with period of time.

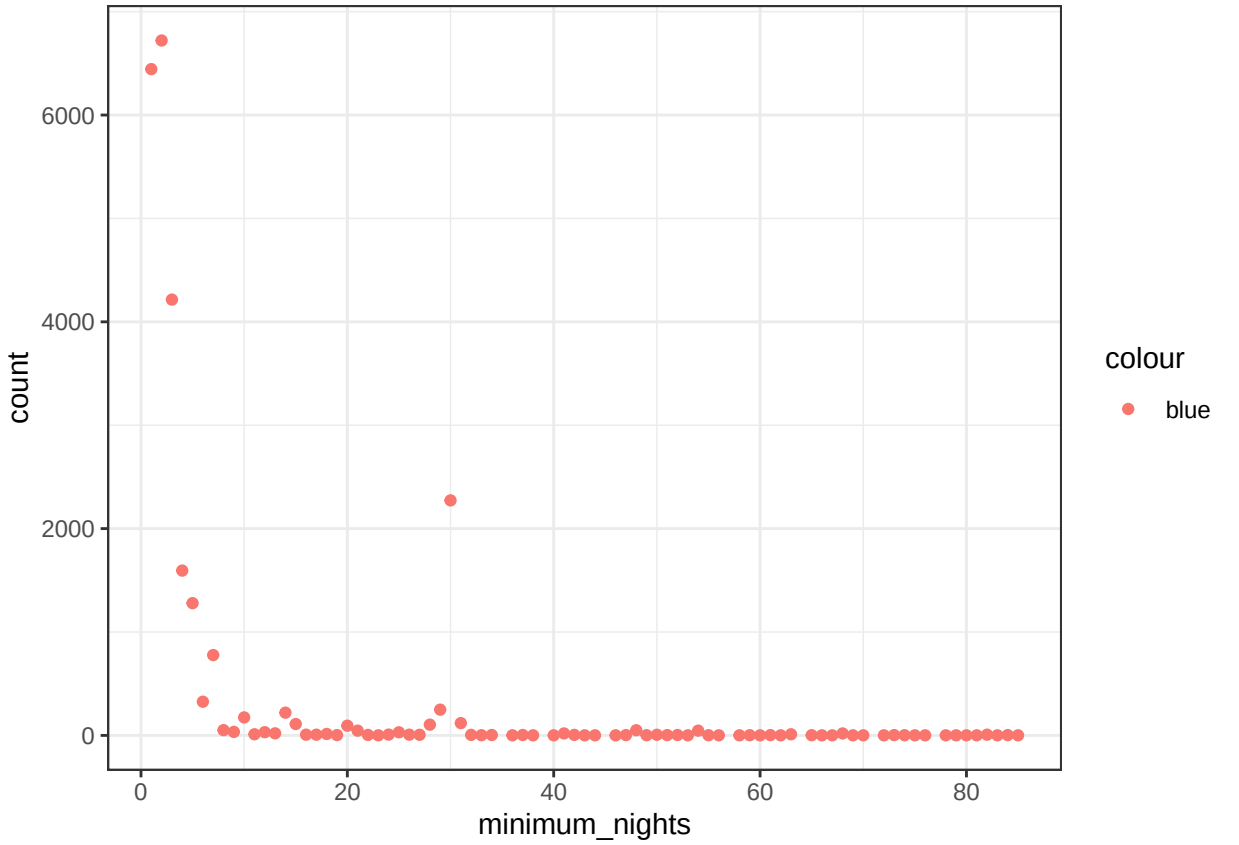
Make new column to category available_365 intro group: 30,90,180,300

```
dfNew$avail_group <- dfNew$availability_365
dfNew$avail_group[dfNew$avail_group<40] = 30
dfNew$avail_group[dfNew$avail_group>=40 & dfNew$avail_group<100] = 90
dfNew$avail_group[dfNew$avail_group>=100 & dfNew$avail_group<200] = 180
dfNew$avail_group[dfNew$avail_group>=200] = 365


dfNew %>%
  group_by(avail_group) %>%
  summarise(count = n()) %>%
  ungroup()
```

```
## # A tibble: 4 x 2
##   avail_group count
##         <dbl> <int>
## 1          30 12915
## 2          90  5012
## 3         180  4853
## 4         365 10029
```

Minimum_night

```
dfNew$minimum_nights <- as.factor(dfNew$minimum_nights)
df_night <- dfNew %>%  filter(availability_365 > 1) %>%
  group_by(minimum_nights) %>%
  summarise(count = n())
df_night <- as.data.frame(df_night)
df_night$minimum_nights <- as.numeric(df_night$minimum_nights)
df_night %>%
  ggplot(aes(y=count,x=minimum_nights, col = "blue")) +
  geom_point() +
  theme_bw()
```



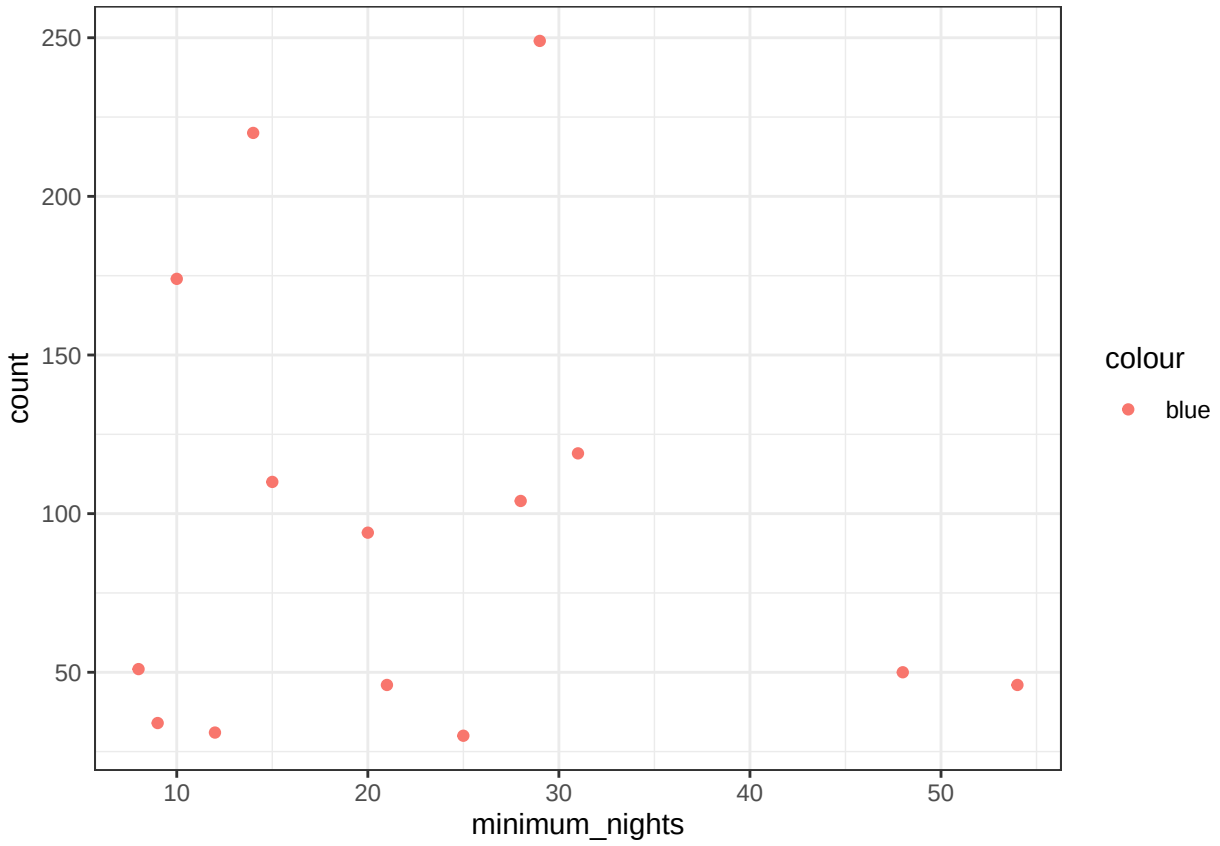
The users input free-style minimum_nights, we can see the numbers are high at 1,2,3,4,5,6,7 and 30. Zoom in to the data different with Minimum_night <=7 and = 30, we can see around 2 weeks, 1 month the data are high, too. So, the user can choose minimum nights are a day, 2 days, 3 days ,a week, 2 weeks, a month, 3 month, 6 month, 12 month.

```
dfNew$minimum_nights <- as.factor(dfNew$minimum_nights)
df_night <- dfNew %>%  filter(availability_365 > 1) %>%
  group_by(minimum_nights) %>%
  summarise(count = n())
df_night %>%
  arrange(desc(count>5))
```

```
## # A tibble: 78 x 2
##    minimum_nights count
##    <fct>          <int>
```

```
##  1 1                6444
##  2 2                6721
##  3 3                4215
##  4 4                1594
##  5 5                1279
##  6 6                 326
##  7 7                 777
##  8 8                  51
##  9 9                  34
## 10 10                174
## # ... with 68 more rows
```

```
df_night <- as.data.frame(df_night)
df_night$minimum_nights <- as.numeric(df_night$minimum_nights)
df_night %>%
  filter(minimum_nights > 7 & minimum_nights != 30 & count > 25) %>%
  ggplot(aes(y=count,x=minimum_nights, col = "blue")) +
  geom_point() +
  theme_bw()
```



Make new column to category minimum_nights intro group: 1,7,15,30,90

```
dfNew$minimum_nights_group <- dfNew$minimum_nights
dfNew$minimum_nights_group[dfNew$minimum_nights_group<4] = 1
```

```
## Warning in Ops.factor(dfNew$minimum_nights_group, 4): '<' not meaningful for
```

```
## factors
```

```
dfNew$minimum_nights_group[dfNew$minimum_nights_group>=4 & dfNew$minimum_nights_group<10] = 7
```

```
## Warning in Ops.factor(dfNew$minimum_nights_group, 4): '>=' not meaningful for
## factors
```

```
## Warning in Ops.factor(dfNew$minimum_nights_group, 10): '<' not meaningful for
## factors
```

```
dfNew$minimum_nights_group[dfNew$minimum_nights_group>=10 & dfNew$minimum_nights_group<20] = 15
```

```
## Warning in Ops.factor(dfNew$minimum_nights_group, 10): '>=' not meaningful for
## factors
```

```
## Warning in Ops.factor(dfNew$minimum_nights_group, 20): '<' not meaningful for
## factors
```

```
dfNew$minimum_nights_group[dfNew$minimum_nights_group>=20 & dfNew$minimum_nights_group<40] = 30
```

```
## Warning in Ops.factor(dfNew$minimum_nights_group, 20): '>=' not meaningful for
## factors
```

```
## Warning in Ops.factor(dfNew$minimum_nights_group, 40): '<' not meaningful for
## factors
```

```
dfNew$minimum_nights_group[dfNew$minimum_nights_group>=40 & dfNew$minimum_nights_group<100] = 90
```

```
## Warning in Ops.factor(dfNew$minimum_nights_group, 40): '>=' not meaningful for
## factors
```

```
## Warning in Ops.factor(dfNew$minimum_nights_group, 100): '<' not meaningful for
## factors
```

```
dfNew$minimum_nights_group[dfNew$minimum_nights_group>=100 & dfNew$minimum_nights_group<300] = 180
```

```
## Warning in Ops.factor(dfNew$minimum_nights_group, 100): '>=' not meaningful for
## factors
```

```
## Warning in Ops.factor(dfNew$minimum_nights_group, 300): '<' not meaningful for
## factors
```

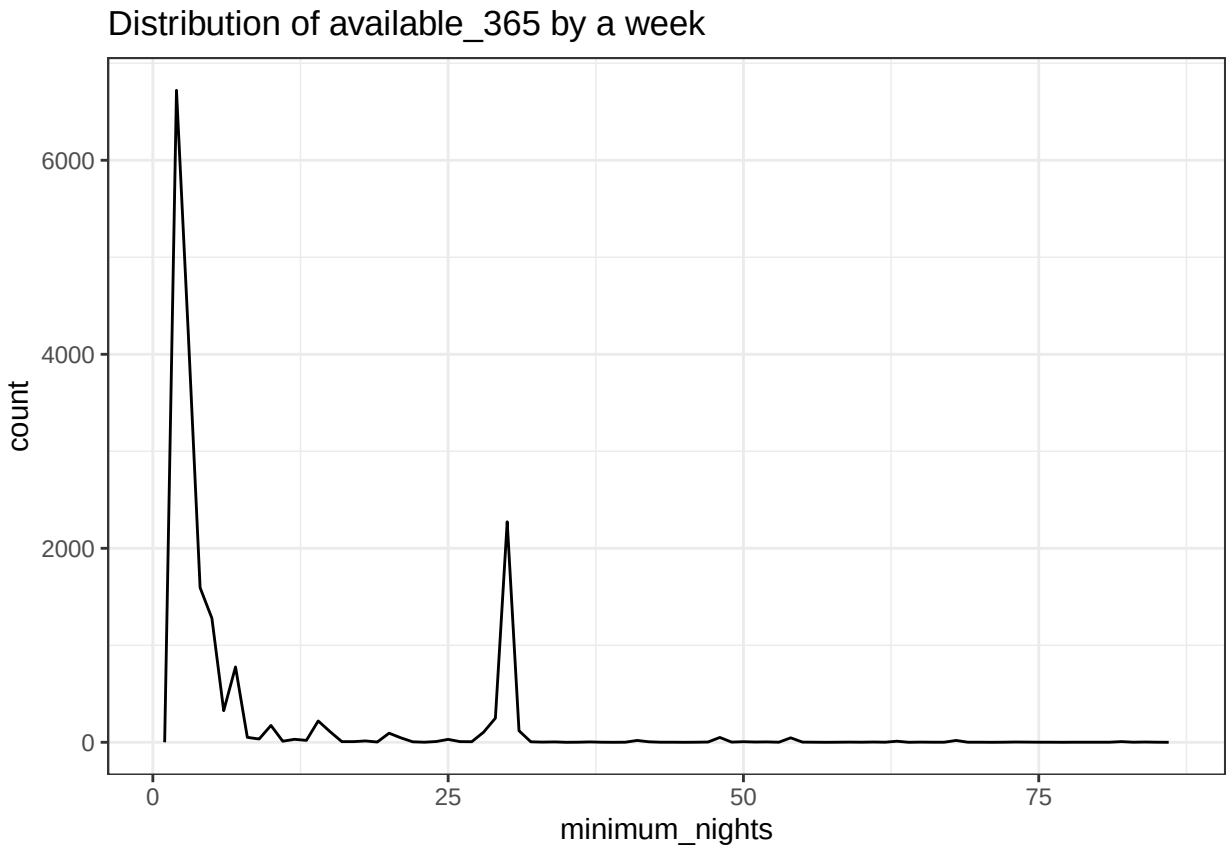
```
dfNew$minimum_nights_group[dfNew$minimum_nights_group>=300 ] = 365
```

```
## Warning in Ops.factor(dfNew$minimum_nights_group, 300): '>=' not meaningful for
## factors
```

```r
dfNew %>%
  group_by(minimum_nights_group) %>%
  summarise(count = n()) %>%
  ungroup()
```

```
## # A tibble: 85 x 2
##    minimum_nights_group count
##    <fct>                <int>
##  1 1                     8742
##  2 2                     8301
##  3 3                     5415
##  4 4                     2168
##  5 5                     1876
##  6 6                      464
##  7 7                     1225
##  8 8                       76
##  9 9                       47
## 10 10                     281
## # ... with 75 more rows
```

```r
dfNew$availability_365 <- as.numeric(dfNew$availability_365)
dfNew$minimum_nights <- as.numeric(dfNew$minimum_nights)
dfNew %>%  filter((availability_365 > 1 & minimum_nights > 1) ) %>%
  ggplot(aes(x=minimum_nights)) +
  geom_freqpoly(binwidth = 1)+
  labs(title = "Distribution of available_365 by a week") +
  theme_bw()
```

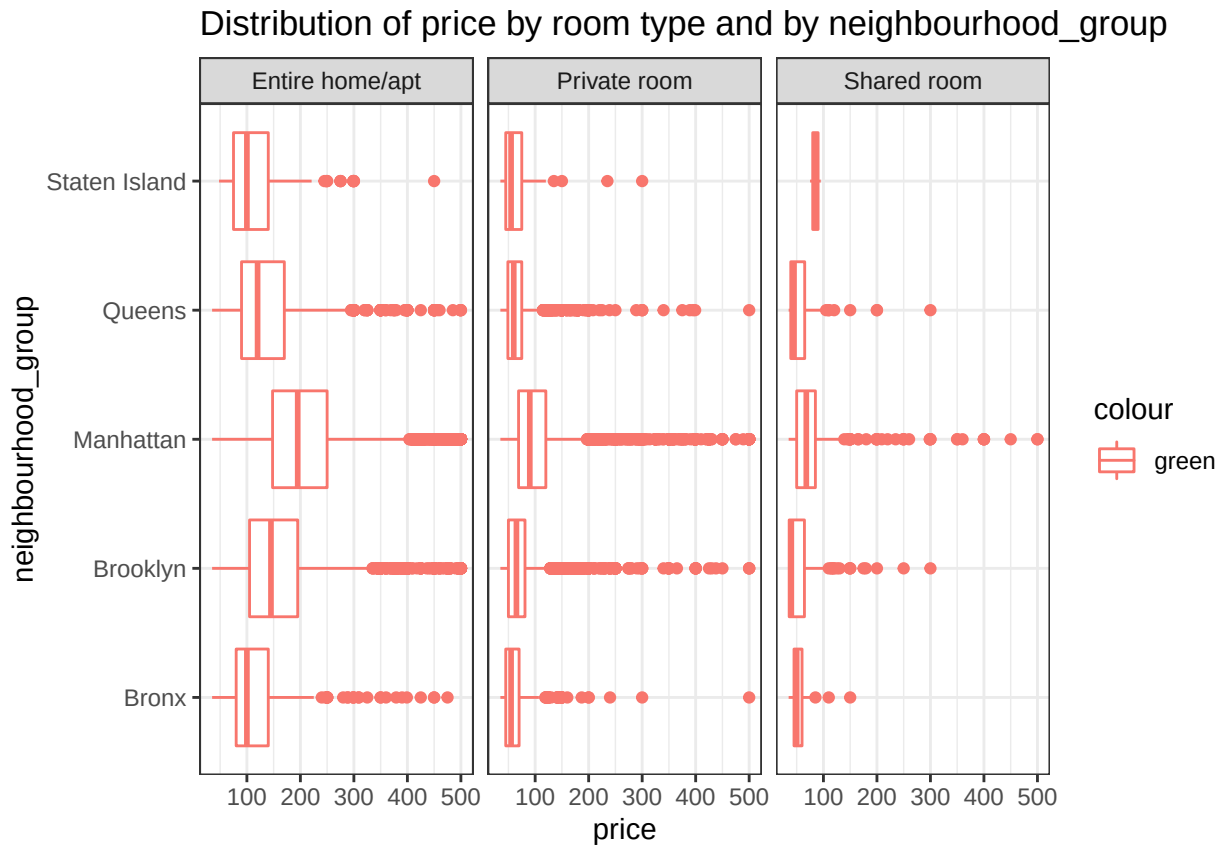



```
dfNew$availability_365 <- as.numeric(dfNew$availability_365)
dfNew %>% filter(availability_365 > 2 & availability_365 < 364  ) %>%
  ggplot(aes(x=minimum_nights,y=price,col=room_type)) +
  geom_point() +
  coord_flip() +
  facet_wrap(~neighbourhood_group) +
  theme_bw() +
  labs(title = "Relationship between price and minimum_nights by room type and by neighborhood group",su
```

As the plot, we can see : About neighborhood : that most of observations in Manhattan. About room_type: - Most of observation in Entire home/apt - Price of entire home are larger than private room About price: - Mininum_nights increase, prices descrease About minimum nights: - Most mininum-nights are <7 and = 30 - Minimum_nights in Manhantan are longer than the others

Price

```
dfNew %>%
  ggplot(aes(x=neighbourhood_group,y=price, col = "green")) +
  geom_boxplot() +
  coord_flip() +
  facet_wrap(~room_type) +
  theme_bw() +
  labs(title = "Distribution of price by room type and by neighbourhood_group")
```



Price depends on room type and neighourhood_group is so clear. Entire home: price from 75 to 250. Private room: price from 50 to 125 Shared room: price from 50 to 75 In Manhattan, there are the prices that larger than 250 usd.

Shared room with the price > 100, no availability_365, host listing only 1 time and no last review means the sections are inactive, so I remove from dataset.

```
dfRemove <- dfNew %>%
  filter(room_type == "Shared room" & price > 100 & availability_365 == 0 & is.na(last_review))
dfNew <- setdiff(dfNew, dfRemove)
```

```
dfNew %>%
  filter(room_type == "Shared room" & price == 500)
```

```
##         id                                name   host_id host_name
## 1 13995008          Awesome Chinatown Apartment 012455431     Tommy
## 2 32089161 Newly rennovated studio space in UWS 025115746 Stephanie
##   neighbourhood_group   neighbourhood latitude longitude   room_type price
## 1           Manhattan    Little Italy 40.71855 -73.99718 Shared room   500
## 2           Manhattan Upper West Side 40.78757 -73.97624 Shared room   500
##   minimum_nights number_of_reviews last_review reviews_per_month
## 1              1                 0        <NA>                NA
## 2              1                 0        <NA>                NA
##   calculated_host_listings_count availability_365 avail_group
## 1                              1                1          30
## 2                              1               90          90
##   minimum_nights_group
```

```
## 1                    1
## 2                    1
```

This id 08324941 post the price is too high so it is available 364/365 days. I remove that to make the price is more reasonable.

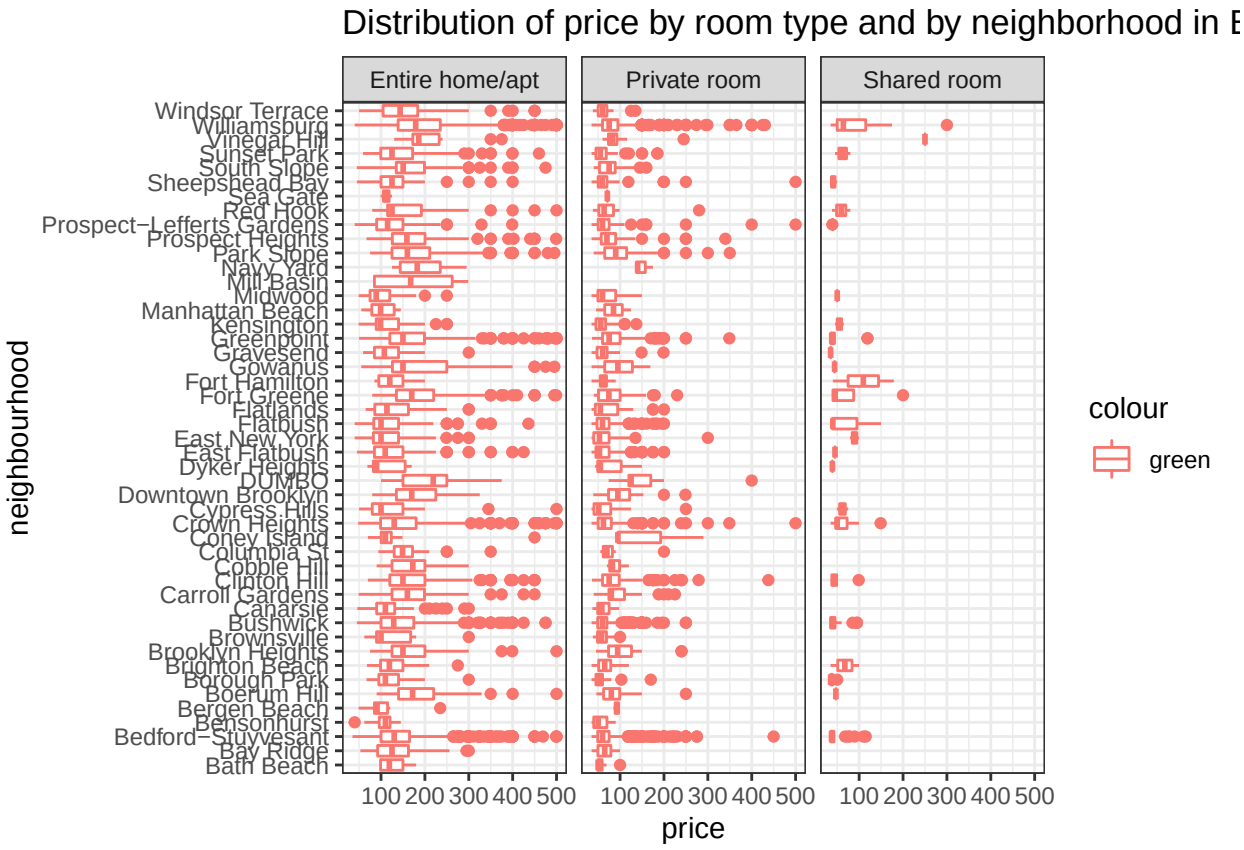
```
dfRemove <- dfNew %>%
  filter(room_type == "Shared room" & price > 100 & neighbourhood_group == "Bronx")
dfNew <- setdiff(dfNew, dfRemove)
```

Shared room with the price > 100, no availability_365 with name Fraud and last review at the Jan, so I remove from dataset.

```
dfRemove <- dfNew %>%
  filter(room_type == "Shared room" & price > 100 & availability_365 == 0)
dfNew <- setdiff(dfNew, dfRemove)
```

```
dfNew %>%
  filter(neighbourhood_group == "Manhattan") %>%
  ggplot(aes(x=neighbourhood,y=price, col = "green")) +
  geom_boxplot() +
  coord_flip() +
  facet_wrap(~room_type) +
  theme_bw() +
  labs(title = "Distribution of price by room type and by neighborhood in Manhattan")
```

```
dfNew %>%
  filter(neighbourhood_group == "Brooklyn") %>%
  ggplot(aes(x=neighbourhood,y=price, col = "green")) +
  geom_boxplot() +
  coord_flip() +
  facet_wrap(~room_type) +
  theme_bw() +
  labs(title = "Distribution of price by room type and by neighborhood in Brooklyn")
```




```
dfNew %>%
  filter(room_type == "Entire home/apt" & price > 300 & host_id == 107434423) %>%
  summary()
```

```
##       id
##  Length:116
##  Class :character
##  Mode  :character
##
##
##
##
##                                                                    name
##  Bright 1BR near Times Sq w/ Indoor pool, Doorman by Blueground       :  2
##  2BR in Hell's Kitchen w/ Gym + Pool close to the subway by Blueground:  1
##  Airy East Village 1BR w/ Doorman, Gym, near NYU by Blueground        :  1
```

```
##  Airy Times Sq 1BR w/ Indoor pool, Gym + Doorman by Blueground        :  1
##  Ample SoHo 2BR w/ W/D, View of Kemare Sq., near Subway, by Blueground:  1
##  Beautiful Chelsea 1BR w/ Balcony, Doorman + Gym by Blueground        :  1
##  (Other)                                                              :109
##       host_id                         host_name      neighbourhood_group
##  107434423:116   Blueground              :116   Bronx        :  0
##  000002571:  0   -TheQueensCornerLot     :  0   Brooklyn     :  1
##  000002787:  0   'Cil                    :  0   Manhattan    :115
##  000002845:  0   (Email hidden by Airbnb):  0   Queens       :  0
##  000002881:  0   (Mary) Haiy             :  0   Staten Island:  0
##  000003151:  0   @ Art House Monique     :  0
##  (Other)  :  0   (Other)                 :  0
##           neighbourhood    latitude       longitude                room_type
##  Chelsea          :22    Min.   :40.70   Min.   :-74.01   Entire home/apt:116
##  Theater District:17    1st Qu.:40.72   1st Qu.:-74.00   Private room   :  0
##  Tribeca          :14    Median :40.74   Median :-73.99   Shared room    :  0
##  West Village     :11    Mean   :40.74   Mean   :-73.99
##  East Village     : 9    3rd Qu.:40.76   3rd Qu.:-73.99
##  Upper West Side  : 9    Max.   :40.79   Max.   :-73.95
##  (Other)          :34
##      price       minimum_nights  number_of_reviews  last_review
##  Min.   :302.0   Min.   :30.00   Min.   :0.00000   Min.   :2019-01-12
##  1st Qu.:314.0   1st Qu.:30.00   1st Qu.:0.00000   1st Qu.:2019-01-24
##  Median :329.5   Median :30.00   Median :0.00000   Median :2019-02-05
##  Mean   :344.2   Mean   :30.83   Mean   :0.09483   Mean   :2019-02-17
##  3rd Qu.:362.8   3rd Qu.:30.00   3rd Qu.:0.00000   3rd Qu.:2019-03-07
##  Max.   :481.0   Max.   :54.00   Max.   :1.00000   Max.   :2019-04-24
##                                                    NA's   :105
##  reviews_per_month calculated_host_listings_count availability_365
##  Min.   :0.1700    Min.   :232                    Min.   :  2.0
##  1st Qu.:0.1800    1st Qu.:232                    1st Qu.:250.0
##  Median :0.1900    Median :232                    Median :311.0
##  Mean   :0.2282    Mean   :232                    Mean   :272.1
##  3rd Qu.:0.2450    3rd Qu.:232                    3rd Qu.:333.0
##  Max.   :0.3900    Max.   :232                    Max.   :363.0
##  NA's   :105
##   avail_group    minimum_nights_group
##  Min.   : 30.0   30     :112
##  1st Qu.:365.0   90     :  4
##  Median :365.0   1      :  0
##  Mean   :321.4   2      :  0
##  3rd Qu.:365.0   3      :  0
##  Max.   :365.0   4      :  0
##                  (Other):  0
```

Hostname Blueground with a lot of Entire home/apartment with price is larger than 300 only in Manhattan with minimum_night more than 1 month.

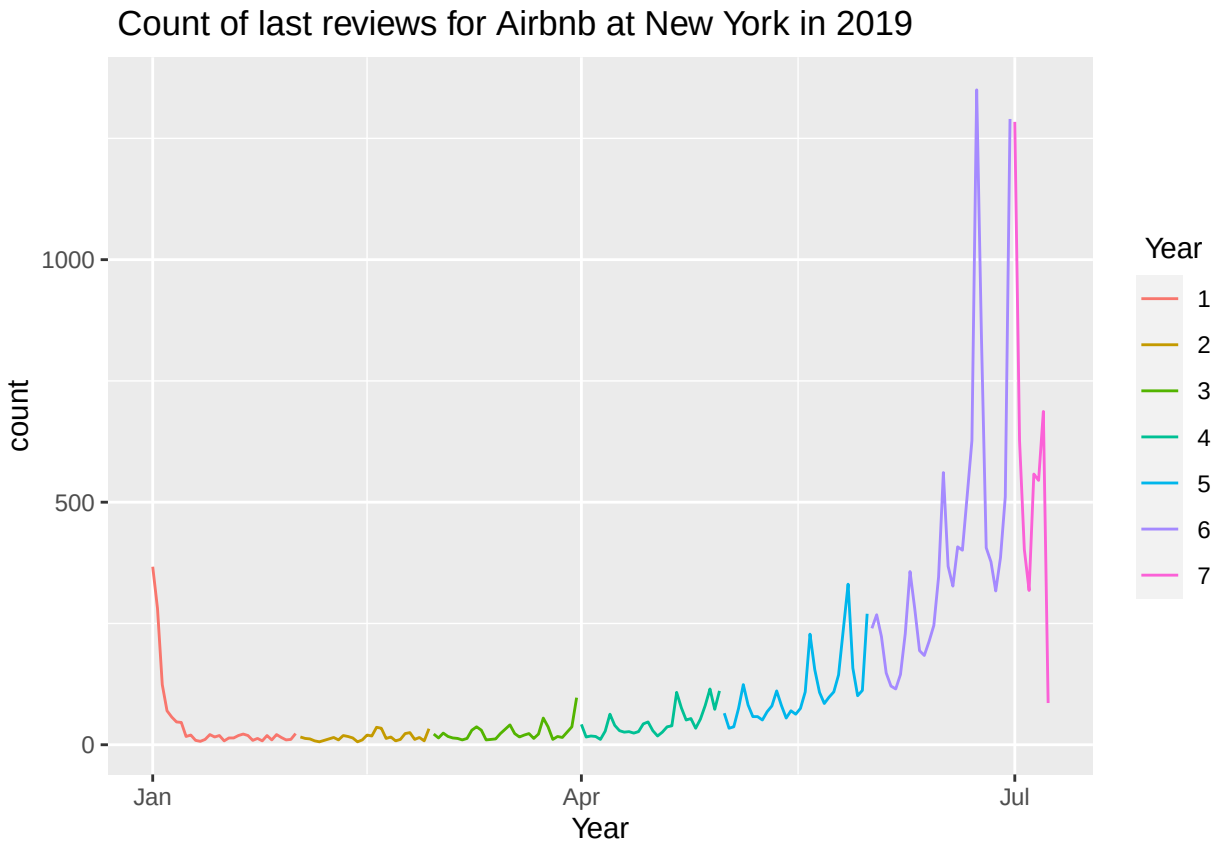
We can see this, because Airbnb give free style input so minimum night, but we can see some period of time. Entirehome: 1 days, 1 week, 2 weeks, 1 month, 2 months, 3 months or long term rental Private room: 1 days, 1 week, 2 weeks, 1 month,(2 months, 3 months is not much) or medium term rental Shared room: short term rental

Most of the post in Mahattan, with minimum nights are longer than other neighborhood group.

Last review

```
dfNew %>%
  ggplot(aes(x=as.Date(last_review), col = as.factor(month(last_review)), group = month(last_review))) +
  labs( x = "Year", col = " Year", title = " Count of last reviews for Airbnb at New York in 2019")
```

```
## Warning: Removed 8963 rows containing non-finite values (stat_count).
```



The lastest last review is 2019-07-08, so there are no more review after this day. As the routine, last review happend in January.

Correlation between numeric variables

We can see that there aren't correlation between all numeric variables : price, minimum_nights, review_per_month, calculated_host_lsiting_count, availability_365.

```
mydata <- dfNew[, c(10,11,14,15,16)]
library(PerformanceAnalytics)
```

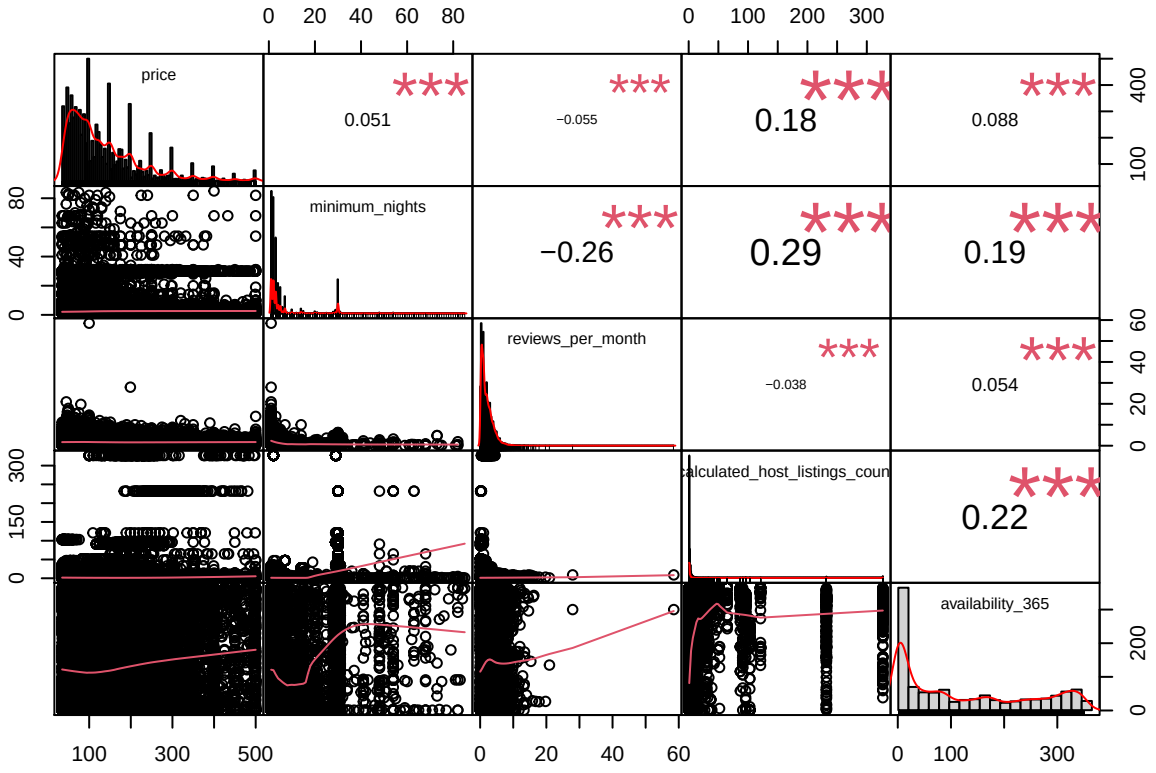
```
## Loading required package: xts
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
##
##     first, last

##
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':
##
##     legend
```

```
chart.Correlation(mydata)
```



Conclusion:

We can use neighbourhood_group, neighbourhood, room_type, minimum_nights_group to train model predict prices. After that we can compare predict price with true price and availability_365 to see How the different of price affect availability_365.