Abnb_Istanbul_Price_Prediction

Pritesh

2/23/2020

```
knitr::opts_chunk$set(echo = TRUE)
library(data.table)

## Warning: package 'data.table' was built under R version 3.6.2

library(fpp)

## Loading required package: forecast

## Warning: package 'forecast' was built under R version 3.6.2

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

## Loading required package: fma

## Warning: package 'fma' was built under R version 3.6.2

## Loading required package: expsmooth

## Loading required package: lmtest

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.6.2

##
## Attaching package: 'zoo'

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

## Loading required package: tseries

library(fpp2)

## Loading required package: ggplot2

##
## Attaching package: 'fpp2'

## The following objects are masked from 'package:fpp':
##
```

```
##     ausair, ausbeer, austa, austourists, debitcards, departures,
##     elecequip, euretail, guinearice, oil, sunspotarea, usmelec
```

```r
library(cowplot)
```

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

```
## 
## ********************************************************
```

```
## Note: As of version 1.0.0, cowplot does not change the
```

```
##   default ggplot2 theme anymore. To recover the previous
```

```
##   behavior, execute:
##   theme_set(theme_cowplot())
```

```
## ********************************************************
```

```r
library(tidyverse)
```

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

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

```
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## v purrr   0.3.3
```

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

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

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

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

```
## -- Conflicts -------------------------------------------------------------
---------- tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()
```

```r
library(psych)
```

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

```
## 
## Attaching package: 'psych'
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.6.2
```

```r
library(dplyr)
library(corrplot)
```

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

## corrplot 0.84 loaded
```

```r
library(GGally)
```

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

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

##
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa

## The following object is masked from 'package:fma':
##
##     pigs
```

```r
AirbnbIstanbul <-
read.csv("C:/Pritesh/Rutgers/Courses/Projects/MVA/Dataset/AirbnbIstanbul.csv"
, stringsAsFactors=FALSE)
Istanbul <- copy(AirbnbIstanbul)
class(Istanbul)
```

```
## [1] "data.frame"
```

```r
setDT(Istanbul)

# data exploration and cleansing #
str(Istanbul) ## to check data type of each var.
```

```
## Classes 'data.table' and 'data.frame':   16251 obs. of  16 variables:
##  $ id                            : int  4826 20815 25436 27271 28277 28308
28318 29241 30697 33368 ...
##  $ name                          : chr  "The Place" "The Bosphorus from
The Comfy Hill" "House for vacation rental furnutare" "LOVELY APT. IN PERFECT
LOCATION" ...
```

```
##  $ host_id                       : int  6603 78838 105823 117026 121607
121695 121721 125742 132137 135136 ...
##  $ host_name                     : chr  "Kaan" "GÃ¼lder" "Yesim" "Mutlu"
...
##  $ neighbourhood_group           : logi  NA NA NA NA NA NA ...
##  $ neighbourhood                 : chr  "Uskudar" "Besiktas" "Besiktas"
"Beyoglu" ...
##  $ latitude                      : num  41.1 41.1 41.1 41 41 ...
##  $ longitude                     : num  29.1 29 29 29 29 ...
##  $ room_type                     : chr  "Entire home/apt" "Entire
home/apt" "Entire home/apt" "Entire home/apt" ...
##  $ price                         : int  554 100 211 237 591 237 633 264
596 295 ...
##  $ minimum_nights                : int  1 30 21 5 3 1 3 3 1 2 ...
##  $ number_of_reviews             : int  1 41 0 2 0 0 0 0 1 1 ...
##  $ last_review                   : chr  "2009-06-01" "2018-11-07" ""
"2018-05-04" ...
##  $ reviews_per_month             : num  0.01 0.38 NA 0.04 NA NA NA NA 0.01
0.02 ...
##  $ calculated_host_listings_count: int  1 2 1 1 13 1 1 1 1 2 ...
##  $ availability_365              : int  365 49 83 228 356 365 365 365 365
232 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
grep('NA',Istanbul) ## indicates NA values are there in 2nd, 5th and 14th
column
```

```
## [1]  2  5 14
```

```
# i.e. name, neighbourhood_group and reviews_per_month have NA values
head(Istanbul,10)
```

```
##        id                                name host_id host_name
##  1:  4826                           The Place    6603      Kaan
##  2: 20815   The Bosphorus from The Comfy Hill   78838   GÃ¼lder
##  3: 25436 House for vacation rental furnutare  105823     Yesim
##  4: 27271      LOVELY APT. IN PERFECT LOCATION  117026     Mutlu
##  5: 28277       Duplex Apartment with Terrace  121607      Alen
##  6: 28308       Great apartment in Cihangir...  121695   Mustafa
##  7: 28318      Cosy home overlooking Bosphorus  121721     Aydin
##  8: 29241     â†ª Istanbul, Your second house  125742    Åževki
##  9: 30697             nice home in popular area  132137       Nan
## 10: 33368   Deluxe double bedroom @ Nisantasi  135136     Ozlem
##     neighbourhood_group neighbourhood latitude longitude       room_type
price
##  1:                  NA       Uskudar 41.05650  29.05367 Entire home/apt
554
##  2:                  NA      Besiktas 41.06984  29.04545 Entire home/apt
100
##  3:                  NA      Besiktas 41.07731  29.03891 Entire home/apt
211
```

```
##  4:                  NA       Beyoglu 41.03220  28.98216 Entire home/apt
237
##  5:                  NA         Sisli 41.04471  28.98567 Entire home/apt
591
##  6:                  NA       Beyoglu 41.03105  28.98297 Entire home/apt
237
##  7:                  NA       Sariyer 41.09048  29.05559 Entire home/apt
633
##  8:                  NA       Beyoglu 41.04844  28.95254    Private room
264
##  9:                  NA       Beyoglu 41.03350  28.97626    Private room
596
## 10:                  NA         Sisli 41.05382  28.99739    Private room
295
##     minimum_nights number_of_reviews last_review reviews_per_month
##  1:              1                 1  2009-06-01              0.01
##  2:             30                41  2018-11-07              0.38
##  3:             21                 0                            NA
##  4:              5                 2  2018-05-04              0.04
##  5:              3                 0                            NA
##  6:              1                 0                            NA
##  7:              3                 0                            NA
##  8:              3                 0                            NA
##  9:              1                 1  2010-06-14              0.01
## 10:              2                 1  2014-10-21              0.02
##     calculated_host_listings_count availability_365
##  1:                              1              365
##  2:                              2               49
##  3:                              1               83
##  4:                              1              228
##  5:                             13              356
##  6:                              1              365
##  7:                              1              365
##  8:                              1              365
##  9:                              1              365
## 10:                              2              232
```

```
dim(Istanbul) # 16251 obs. and 16 vars
```

```
## [1] 16251    16
```

```
summary(Istanbul) ## summarized view of all the feature/vars
```

```
##        id                name                host_id             host_name
##  Min.   :    4826   Length:16251       Min.   :     6603   Length:16251
##  1st Qu.: 8500978   Class :character   1st Qu.: 17882300   Class
:character
##  Median :21619750   Mode  :character   Median : 52107399   Mode
:character
##  Mean   :18856396                      Mean   : 88887056
##  3rd Qu.:28702192                      3rd Qu.:168134520
##  Max.   :32457561                      Max.   :243734065  
##                                                            
##  neighbourhood_group neighbourhood         latitude       longitude    
##  Mode:logical        Length:16251       Min.   :40.81   Min.   :28.03  
##  NA's:16251          Class :character   1st Qu.:41.00   1st Qu.:28.97  
##                      Mode  :character   Median :41.03   Median :28.98  
##                                         Mean   :41.03   Mean   :28.99  
##                                         3rd Qu.:41.05   3rd Qu.:29.02  
##                                         Max.   :41.41   Max.   :29.91  
##                                                                        
##   room_type             price         minimum_nights    number_of_reviews
##  Length:16251       Min.   :    0.0   Min.   :   1.000   Min.   :  0.000  
##  Class :character   1st Qu.:  105.0   1st Qu.:   1.000   1st Qu.:  0.000  
##  Mode  :character   Median :  190.0   Median :   1.000   Median :  0.000  
##                     Mean   :  354.7   Mean   :   4.693   Mean   :  7.187  
##                     3rd Qu.:  327.0   3rd Qu.:   2.000   3rd Qu.:  4.000  
##                     Max.   :59561.0   Max.   :1125.000   Max.   :307.000  
##                                                                           
##  last_review        reviews_per_month calculated_host_listings_count
##  Length:16251       Min.   : 0.010    Min.   : 1.000                
##  Class :character   1st Qu.: 0.180    1st Qu.: 1.000                
##  Mode  :character   Median : 0.520    Median : 1.000                
##                     Mean   : 0.915    Mean   : 4.104                
##                     3rd Qu.: 1.190    3rd Qu.: 4.000                
##                     Max.   :12.000    Max.   :77.000                
##                     NA's   :8484                                    
##  availability_365
##  Min.   :  0.0   
##  1st Qu.:101.0   
##  Median :340.0   
##  Mean   :249.5   
##  3rd Qu.:365.0   
##  Max.   :365.0   
##
```

```
unique(Istanbul$room_type) ## 3 unique room types
```

```
## [1] "Entire home/apt" "Private room"    "Shared room"
```

```
unique(Istanbul$neighbourhood) ## 39 unique neighbourhoods
```

```
##  [1] "Uskudar"       "Besiktas"      "Beyoglu"       "Sisli"        
##  [5] "Sariyer"       "Beykoz"        "Atasehir"      "Fatih"        
##  [9] "Adalar"        "Kadikoy"       "Kagithane"     "Maltepe"      
## [13] "Bakirkoy"      "Esenyurt"      "Basaksehir"    "Kartal"       
## [17] "Gaziosmanpasa" "Bahcelievler"  "Bagcilar"      "Buyukcekmece" 
## [21] "Silivri"       "Beylikduzu"    "Umraniye"      "Sile"         
## [25] "Cekmekoy"      "Sancaktepe"    "Tuzla"         "Pendik"       
## [29] "Sultangazi"    "Eyup"          "Zeytinburnu"   "Kucukcekmece" 
## [33] "Avcilar"       "Gungoren"      "Catalca"       "Bayrampasa"   
## [37] "Esenler"       "Sultanbeyli"   "Arnavutkoy"
## since, I used stringsAsFactors=FALSE while importing the dataset, few of
the columns
## like name, host_name, neighbourhood and room_type belongs to character
data type
## hence, will factor neighbourhood and room_type for now. name and host_name
doesn't seem
## to be much interest for now, hence will leave those.
str(Istanbul)

## Classes 'data.table' and 'data.frame':   16251 obs. of  16 variables:
##  $ id                            : int  4826 20815 25436 27271 28277 28308
28318 29241 30697 33368 ...
##  $ name                          : chr  "The Place" "The Bosphorus from
The Comfy Hill" "House for vacation rental furnutare" "LOVELY APT. IN PERFECT
LOCATION" ...
##  $ host_id                       : int  6603 78838 105823 117026 121607
121695 121721 125742 132137 135136 ...
##  $ host_name                     : chr  "Kaan" "GÃ¼lder" "Yesim" "Mutlu"
...
##  $ neighbourhood_group           : logi  NA NA NA NA NA NA ...
##  $ neighbourhood                 : chr  "Uskudar" "Besiktas" "Besiktas"
"Beyoglu" ...
##  $ latitude                      : num  41.1 41.1 41.1 41 41 ...
##  $ longitude                     : num  29.1 29 29 29 29 ...
##  $ room_type                     : chr  "Entire home/apt" "Entire
home/apt" "Entire home/apt" "Entire home/apt" ...
##  $ price                         : int  554 100 211 237 591 237 633 264
596 295 ...
##  $ minimum_nights                : int  1 30 21 5 3 1 3 3 1 2 ...
##  $ number_of_reviews             : int  1 41 0 2 0 0 0 0 1 1 ...
##  $ last_review                   : chr  "2009-06-01" "2018-11-07" ""
"2018-05-04" ...
##  $ reviews_per_month             : num  0.01 0.38 NA 0.04 NA NA NA NA 0.01
0.02 ...
##  $ calculated_host_listings_count: int  1 2 1 1 13 1 1 1 1 2 ...
##  $ availability_365              : int  365 49 83 228 356 365 365 365 365
232 ...
##  - attr(*, ".internal.selfref")=<externalptr>

Istanbul[,room_type:=factor(room_type)]
Istanbul[,neighbourhood:=factor(neighbourhood)]
Istanbul[,last_review:=as.Date(last_review,'%Y-%m-%d')] ## converting
last_review to date datatype

# datatypes looks better now. hence will see again for NA values
grep ('NA',Istanbul) # 2, 5, 13 and 14 column have NA values

## [1]  2  5 13 14

Istanbul[is.na(neighbourhood_group),NROW(neighbourhood_group)] # entire obs.
is blank, will drop this var
```

```
## [1] 16251
```

```
Istanbul[is.na(last_review),NROW(last_review)] ## there are 8484 NA values
```

```
## [1] 8484
```

```
Istanbul[is.na(reviews_per_month),NROW(reviews_per_month)] ## there are 8484
NA values
```

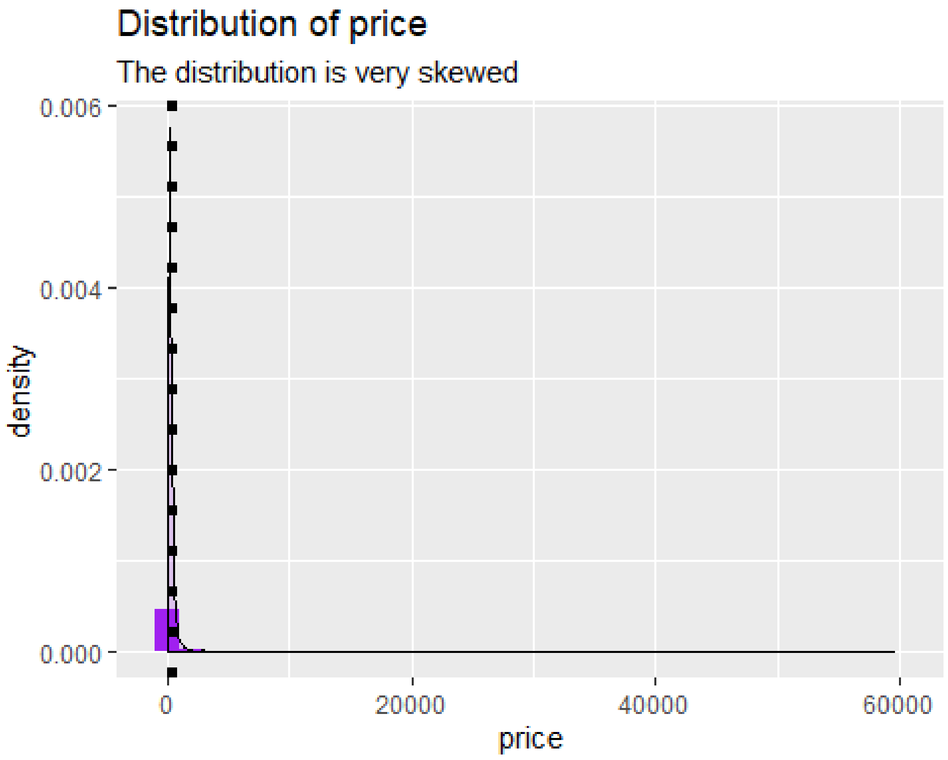
```
## [1] 8484
```

```
Istanbul$neighbourhood_group <- NULL ## removing neighbourhood_group column
Istanbul[is.na(reviews_per_month),reviews_per_month:=0] ## nearly 50% of the
dataset is filled with NA.
# hence we can't simply remove these many rows. Hence imputing with 0 values.

# performing exploratory data analysis #

str(Istanbul)
```

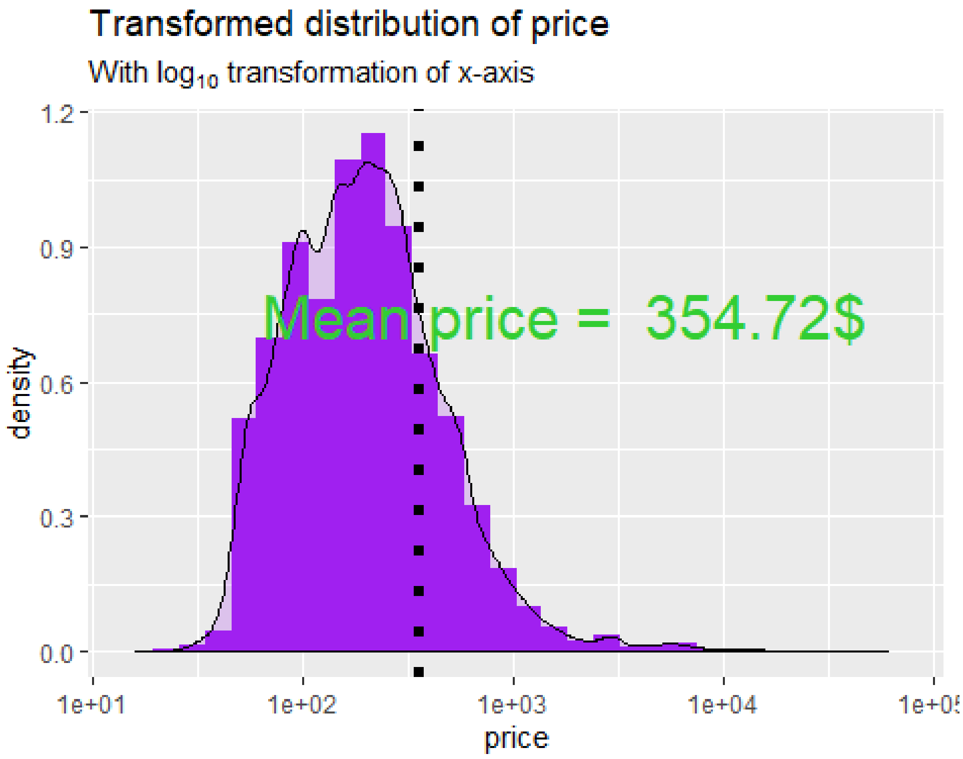
```
## Classes 'data.table' and 'data.frame':   16251 obs. of  15 variables:
##  $ id                            : int  4826 20815 25436 27271 28277 28308
28318 29241 30697 33368 ...
##  $ name                          : chr  "The Place" "The Bosphorus from
The Comfy Hill" "House for vacation rental furnutare" "LOVELY APT. IN PERFECT
LOCATION" ...
##  $ host_id                       : int  6603 78838 105823 117026 121607
121695 121721 125742 132137 135136 ...
##  $ host_name                     : chr  "Kaan" "GÃ¼lder" "Yesim" "Mutlu"
...
##  $ neighbourhood                 : Factor w/ 39 levels
"Adalar","Arnavutkoy",..: 38 10 10 13 33 13 30 13 13 33 ...
##  $ latitude                      : num  41.1 41.1 41.1 41 41 ...
##  $ longitude                     : num  29.1 29 29 29 29 ...
##  $ room_type                     : Factor w/ 3 levels "Entire
home/apt",..: 1 1 1 1 1 1 1 2 2 2 ...
##  $ price                         : int  554 100 211 237 591 237 633 264
596 295 ...
##  $ minimum_nights                : int  1 30 21 5 3 1 3 3 1 2 ...
##  $ number_of_reviews             : int  1 41 0 2 0 0 0 0 1 1 ...
##  $ last_review                   : Date, format: "2009-06-01" "2018-11-07"
...
##  $ reviews_per_month             : num  0.01 0.38 0 0.04 0 0 0 0 0.01 0.02
...
##  $ calculated_host_listings_count: int  1 2 1 1 13 1 1 1 1 2 ...
##  $ availability_365              : int  365 49 83 228 356 365 365 365 365
232 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
dim(Istanbul) # 16251 obs. and 15 vars, with last_review in date format
```

```
## [1] 16251    15

# price looks to be our dependent variable, hence will see the distribution
of price
summary(Istanbul$price)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   105.0   190.0   354.7   327.0 59561.0

ggplot(Istanbul, aes(price)) +
  geom_histogram(bins = 30, aes(y = ..density..), fill = "purple") +
  geom_density(alpha = 0.2, fill = "purple") +
  ggtitle("Distribution of price",
          subtitle = "The distribution is very skewed") +
  theme(axis.title = element_text(), axis.title.x = element_text()) +
  geom_vline(xintercept = round(mean(Istanbul$price), 2), size = 2, linetype
= 3)
```



```
#As distribution is very skewed, performing logarithmic transformation to
gain better insight
ggplot(Istanbul, aes(price)) +
  geom_histogram(bins = 30, aes(y = ..density..), fill = "purple") +
  geom_density(alpha = 0.2, fill = "purple") +
  ggtitle("Transformed distribution of price",
          subtitle = expression("With" ~'log'[10] ~ "transformation of x-
axis")) +
  #theme(axis.title = element_text(), axis.title.x = element_text()) +
  geom_vline(xintercept = round(mean(Istanbul$price), 2), size = 2, linetype
= 3) +
  scale_x_log10() +
  annotate("text", x = 1800, y = 0.75,label = paste("Mean price = ",
paste0(round(mean(Istanbul$price), 2), "$")),
           color =  "#32CD32", size = 8)

## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Removed 3 rows containing non-finite values (stat_bin).

## Warning: Removed 3 rows containing non-finite values (stat_density).
```

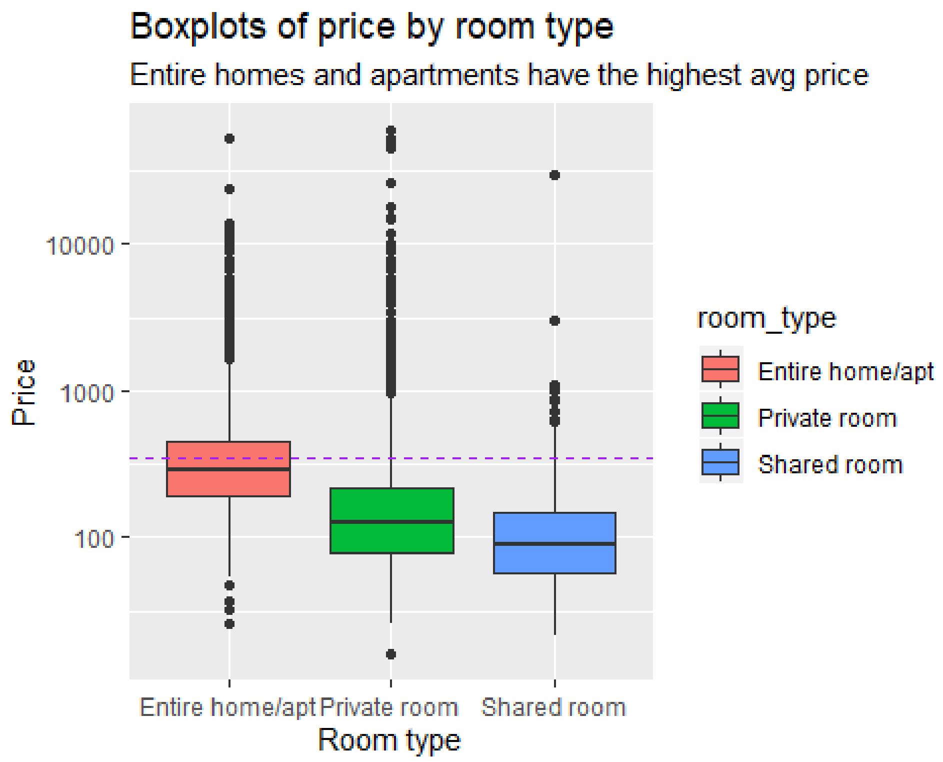


```
#What drives price? Checking Price values with respect to KPIs

#1 relationship between price and room type
describeBy(Istanbul$price,Istanbul$room_type)

##
##  Descriptive statistics by group
## group: Entire home/apt
##    vars    n   mean     sd median trimmed    mad min   max range  skew
kurtosis
## X1    1 7191 425.88 913.75    285  315.58 164.57   0 52728 52728 31.39
1580.85
##       se
## X1 10.78
## ---------------------------------------------------------------
## group: Private room
##    vars    n   mean      sd median trimmed   mad min   max range skew
kurtosis
## X1    1 8565 303.78 1749.27    127   150.8 78.58  16 59561 59545 25.6
769.35
##      se
## X1 18.9
## ---------------------------------------------------------------
## group: Shared room
##    vars   n   mean      sd median trimmed   mad min   max range  skew
kurtosis
## X1    1 495 202.51 1351.03     90   104.3 47.44  21 29786 29765 21.25
461.52
##       se
## X1 60.72
```

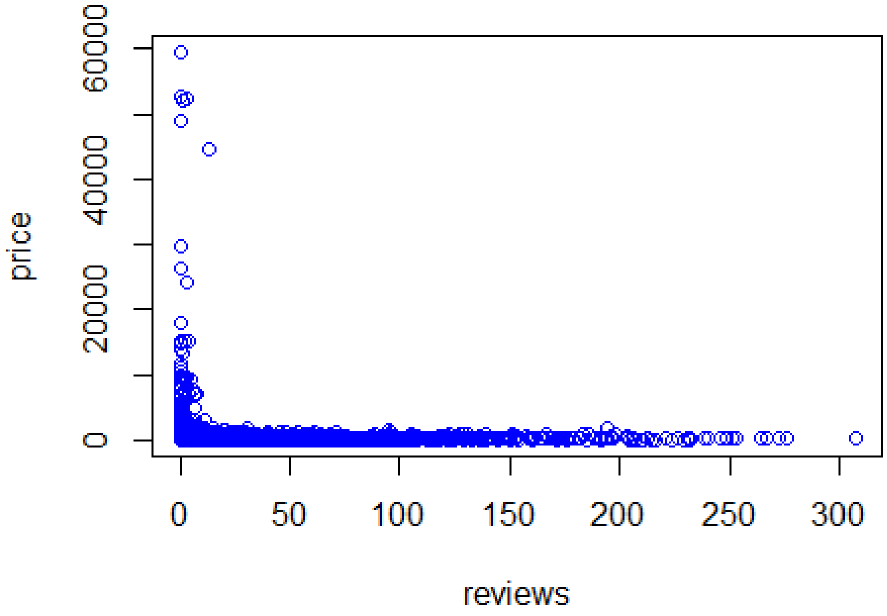
```r
ggplot(Istanbul, aes(x = room_type, y = price)) +
  geom_boxplot(aes(fill = room_type)) + scale_y_log10() +
  xlab("Room type") +
  ylab("Price") +
  ggtitle("Boxplots of price by room type",
          subtitle = "Entire homes and apartments have the highest avg
price") +
  geom_hline(yintercept = mean(Istanbul$price), color = "purple", linetype =
2)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

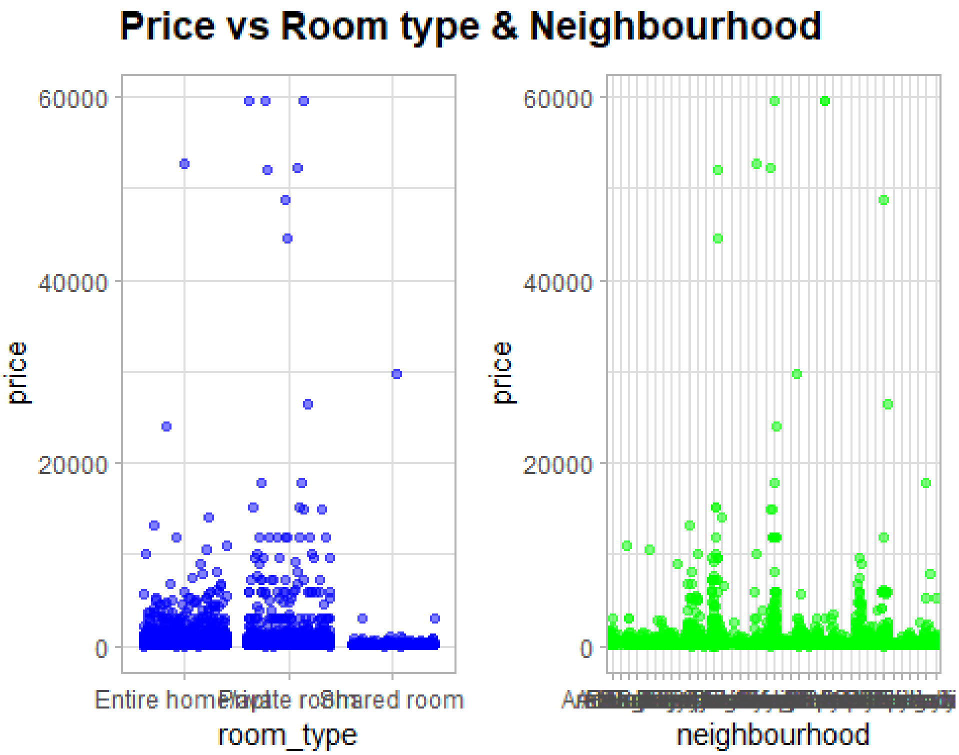
```
## Warning: Removed 3 rows containing non-finite values (stat_boxplot).
```



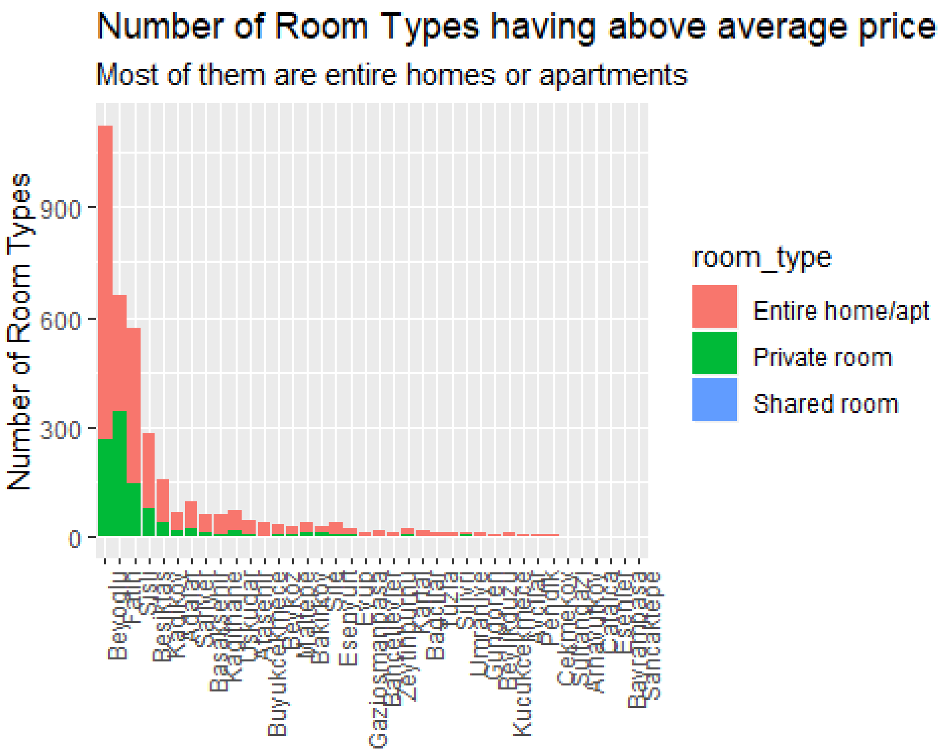
```
# We see that Entire Home/Apartments have the highest avg price. Also the
private room's
# prices are comparable to Entire Home/Apartments price

#2 price vs number of reviews
plot(price ~ number_of_reviews, data=Istanbul,xlab='reviews', ylab =
'price',col='blue')
```



```r
#The most pricy listings have lesser number of reviews

#3 price vs room type and neighbourhoods
#Scatter plot in one screen, Price vs Room type & Neighbourhood
x <- ggplot(Istanbul, aes(room_type, price)) +
  geom_jitter(color = "blue", alpha = 0.5) +
  theme_light()

y <- ggplot(Istanbul, aes(neighbourhood, price)) +
  geom_jitter(color = "green", alpha = 0.5) +
  theme_light()

p <- plot_grid(x, y)
title <- ggdraw() + draw_label("Price vs Room type & Neighbourhood",
fontface='bold')
plot_grid(title, p, ncol=1, rel_heights=c(0.1, 1))
```



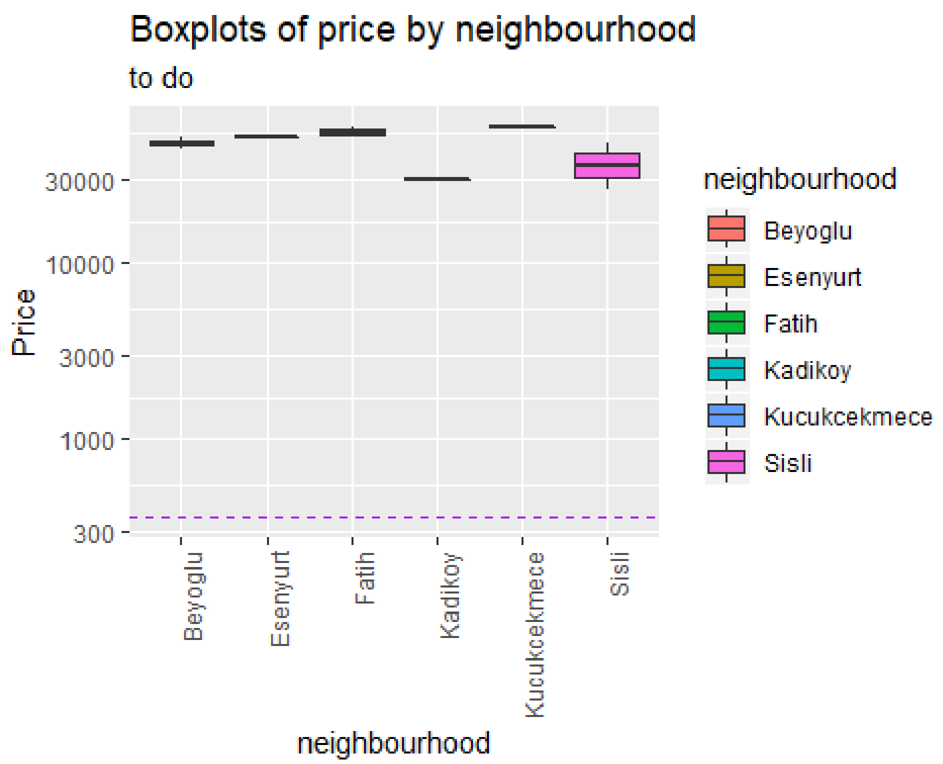
```
# The scatter plot doesn't give clear picture, hence will draw the bar chart
# Above Average Price by Neighourhood Areas and room_type together.
Istanbul %>% filter(price >= mean(price)) %>% group_by(neighbourhood,
room_type) %>% tally %>%
  ggplot(aes(reorder(neighbourhood,desc(n)), n, fill = room_type)) +
  xlab(NULL) +
  ylab("Number of Room Types") +
  ggtitle("Number of Room Types having above average price",
          subtitle = "Most of them are entire homes or apartments") +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



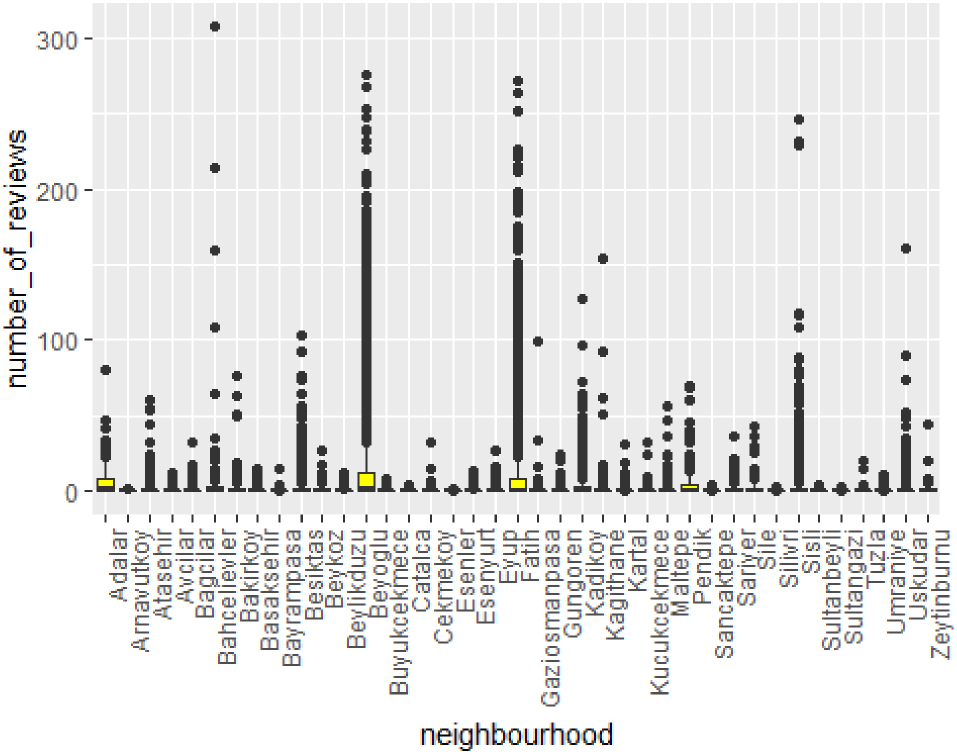
```
# Beyoglu, Sisli and Fatih neighbourhoods have more than the average price
# as well have more number of units than other neighbourhoods.

# Top 10 most priced locations
range(Istanbul$price) ## range of price

## [1]     0 59561

avgNeighbourhood=Istanbul[,avgneighprice:=mean(price),by=neighbourhood]
Istanbul.1 <- avgNeighbourhood[price > avgneighprice]
top10localities <- head(arrange(Istanbul.1,desc(Istanbul.1$price)), n = 10)
top10localities

##          id                                                    name
host_id
## 1  30361326        3 Rooms 1 Living Room - Grand Holiday Istanbul
227944870
## 2  30361470 3 Rooms 1 Living Room Dublex - Grand Holiday Istanbul
227944870
## 3  31974054                  Elegance Single Room - Avicenna Hotel
166950259
## 4  22119662                                    Gunluk kiralik daire
161593238
## 5  29257295                              Ä°stanbul town history place
20973637
## 6  19619789                            CoZy room in BeyoÄŸlu/cihangir
36781586
```

```
## 7  20275354    Ä°stanbul un kalbi sisli. Center of istanbul sisli
118695718
## 8   7016669                   Private room in BeyoÄŸlu(nice view)
36781586
## 9  19235485                                                  hmgv
134667922
## 10 13703737                            Room in the center BOMONTI
3162918
##    host_name neighbourhood latitude longitude       room_type price
## 1      Ilhan  Kucukcekmece 41.03740  28.79435    Private room 59561
## 2      Ilhan  Kucukcekmece 41.03841  28.79471    Private room 59561
## 3   Avicenna         Fatih 41.00445  28.97907    Private room 59561
## 4     Leylan      Esenyurt 41.02681  28.62680 Entire home/apt 52728
## 5       Memo         Fatih 41.00850  28.96649    Private room 52243
## 6       Kaan       Beyoglu 41.03015  28.98064    Private room 52000
## 7       Ipek         Sisli 41.05465  28.98111    Private room 48842
## 8       Kaan       Beyoglu 41.03383  28.97151    Private room 44671
## 9      Deniz       Kadikoy 40.99484  29.02976     Shared room 29786
## 10   Cagatay         Sisli 41.05709  28.98525    Private room 26364
##    minimum_nights number_of_reviews last_review reviews_per_month
## 1               1                 0        <NA>              0.00
## 2               1                 0        <NA>              0.00
## 3               1                 0        <NA>              0.00
## 4               5                 0        <NA>              0.00
## 5               2                 3  2018-11-03              0.75
## 6               1                 1  2017-07-21              0.05
## 7               2                 0        <NA>              0.00
## 8               1                13  2016-04-25              0.30
## 9             300                 0        <NA>              0.00
## 10              1                 0        <NA>              0.00
##    calculated_host_listings_count availability_365 avgneighprice
## 1                               3              360     1263.4643
## 2                               3              331     1263.4643
## 3                               4              363      498.9310
## 4                               1                0      403.1296
## 5                               1              359      498.9310
## 6                               2               89      373.1771
## 7                               1                0      342.1759
## 8                               2              363      373.1771
## 9                               1              364      204.3891
## 10                              1              364      342.1759
```

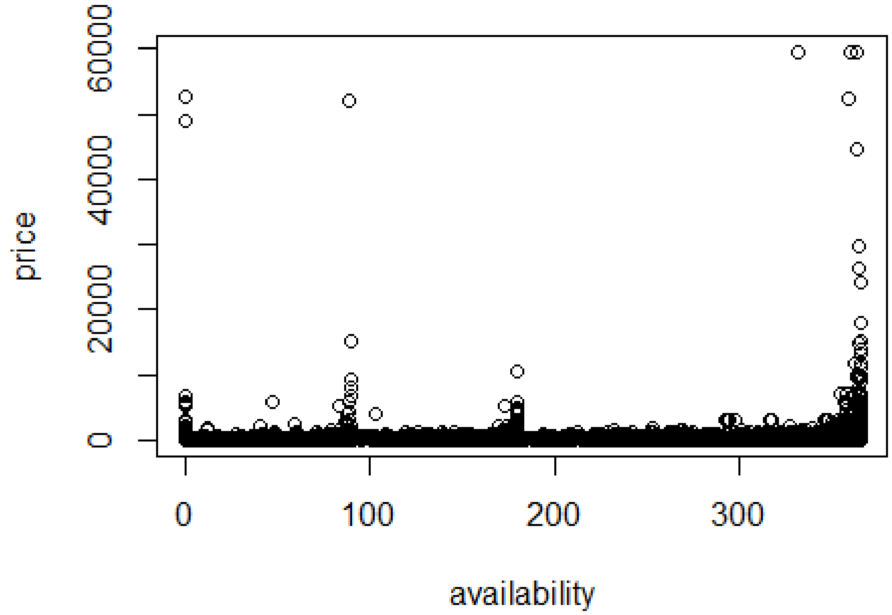
```
ggplot(top10localities, aes(x = neighbourhood, y = price)) +
  geom_boxplot(aes(fill = neighbourhood)) +  scale_y_log10() +
  xlab("neighbourhood") +
  ylab("Price") +
  ggtitle("Boxplots of price by neighbourhood",
          subtitle = "to do") +
  geom_hline(yintercept = mean(Istanbul$price), color = "purple", linetype =
2) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



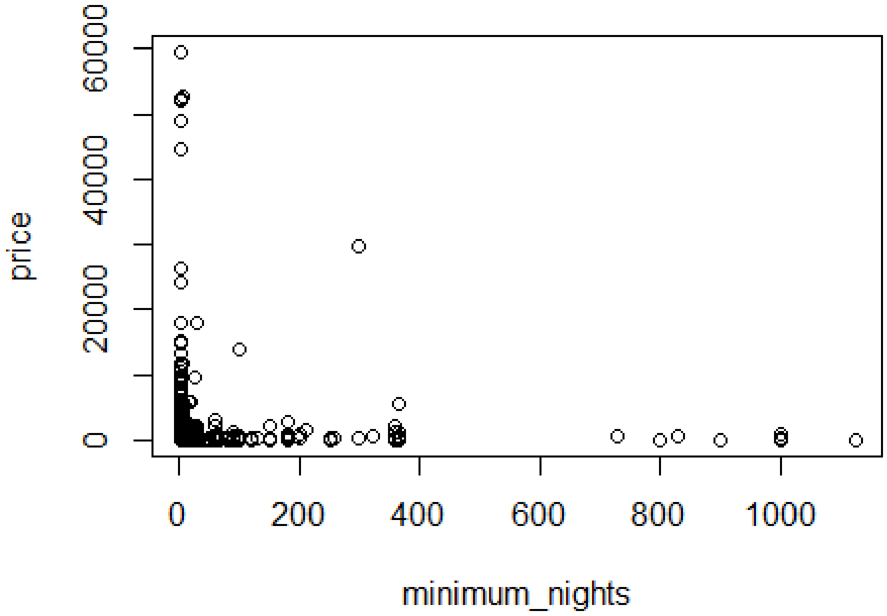
```
#4 no. of reviews and neighbourhood relation
summary(Istanbul$number_of_reviews)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   7.187   4.000 307.000

ggplot(Istanbul,aes(x=neighbourhood,y=number_of_reviews)) +
geom_boxplot(fill='yellow') + theme(axis.text.x = element_text(angle = 90,
hjust = 1))
```



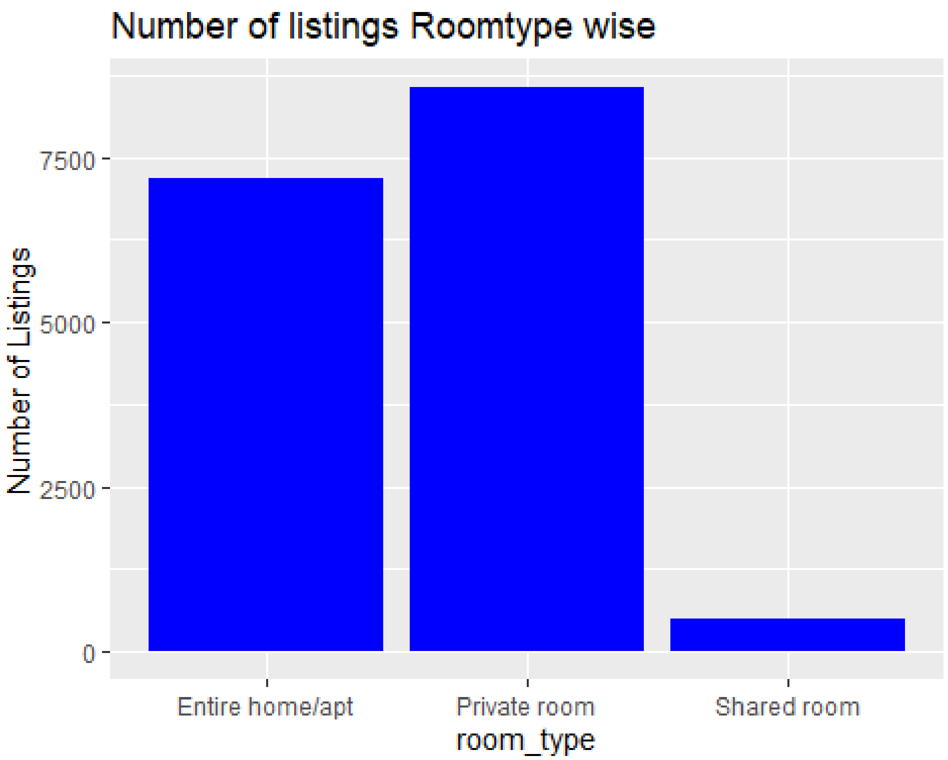
```r
# Top 10 neighbourhoods having most number of reviews, pending
# top10reviews_by_locality <-
head(arrange(Istanbul,desc(Istanbul$number_of_reviews)), n = 10)
# top10reviews_by_locality
# ggplot(top10reviews_by_locality,aes(x=neighbourhood,y=number_of_reviews)) +
geom_boxplot(fill='yellow') + theme(axis.text.x = element_text(angle = 90,
hjust = 1))

#5 price vs availability relation
plot(price ~ availability_365, data=Istanbul,xlab='availability', ylab =
'price')
```



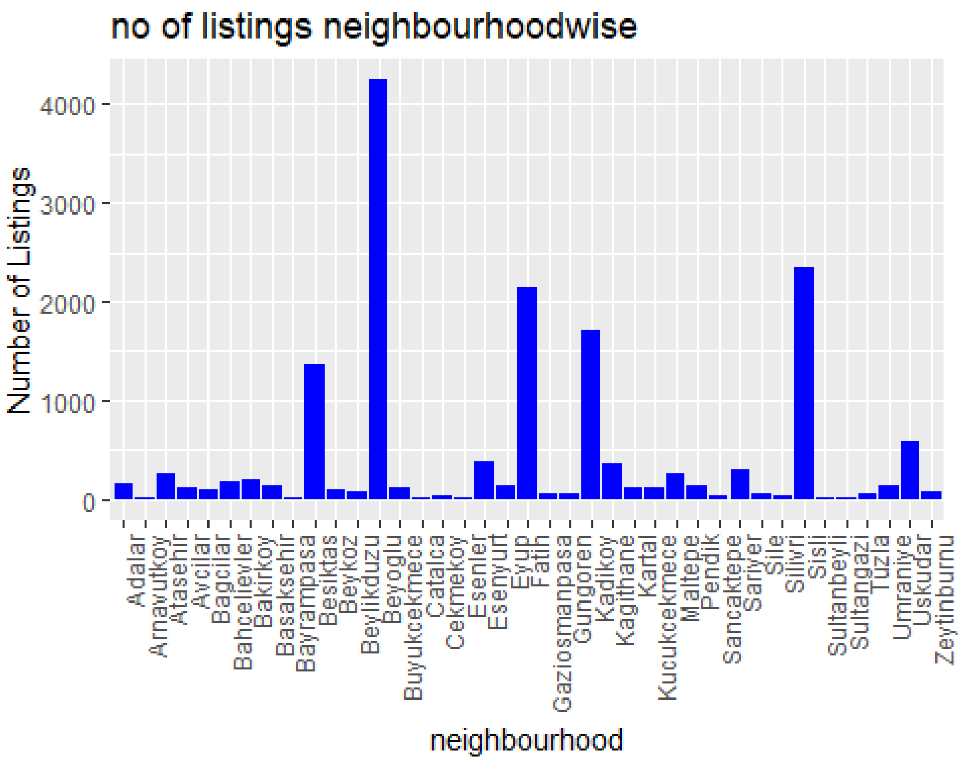
```
#It is hard to see a clear pattern but the most priced listings have either
very few days availability
# or maximum days availability

#6 price vs minimum nights relation
plot(price ~ minimum_nights, data=Istanbul,xlab='minimum_nights', ylab =
'price')
```



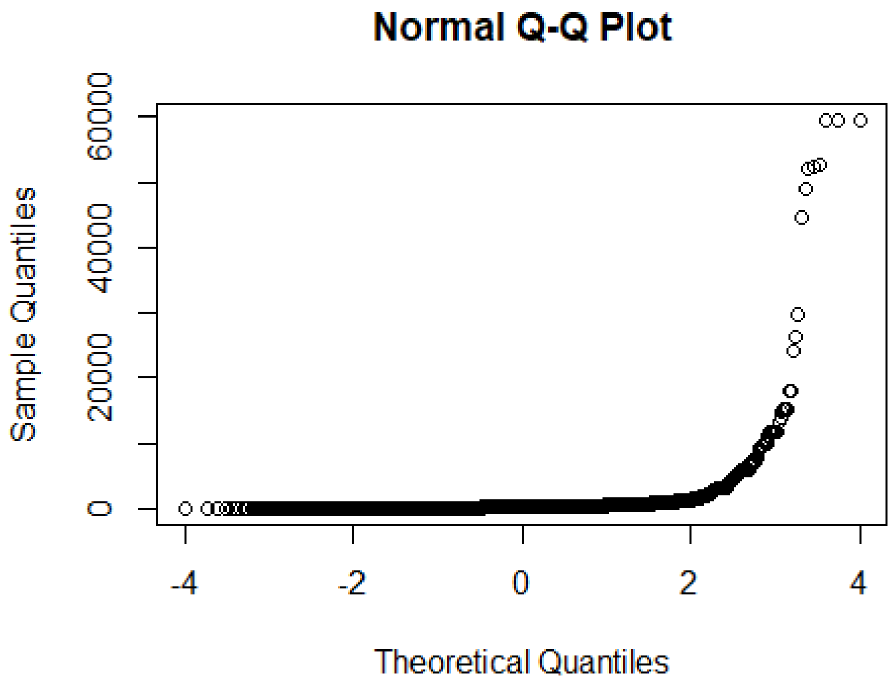
```
#with lesser number of 'min no of nights' , Prices are high and Prices
decrease with increase in Min no of nights

#7 listing vs room type relation
#no of listings vs room type
ggplot(Istanbul,aes(x=room_type)) + geom_bar(fill = 'blue')+
  ylab("Number of Listings") +
  ggtitle("Number of listings Roomtype wise")
```



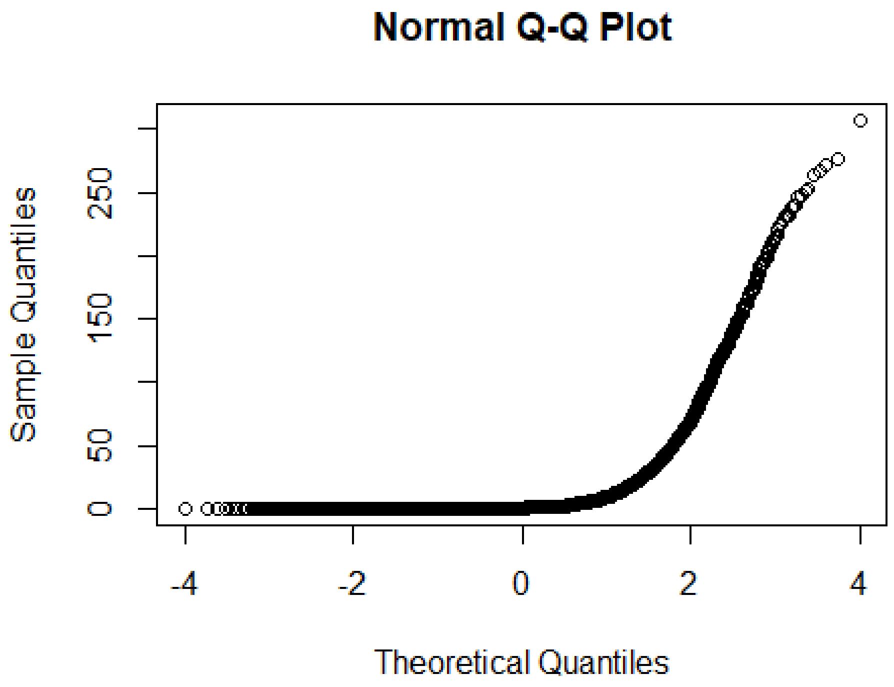
```
#Private rooms are more in number

#8 no of listings neighbourhoodwise
ggplot(Istanbul,aes(x=neighbourhood)) + geom_bar(fill = 'blue') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ylab("Number of Listings") +
  ggtitle("no of listings neighbourhoodwise")
```



```
# Beyoglu, Sisli and Fatih have most number of listings


# checking correlation between all the variables
Istanbul.2 <- Istanbul[,c(6,7,9,10,11,13,14,15)] ## filtering dataset
containing only numberical data
## qqnorm plot ## indicating that the variables are not normalized
q1 = qqnorm(Istanbul.2$price)
```



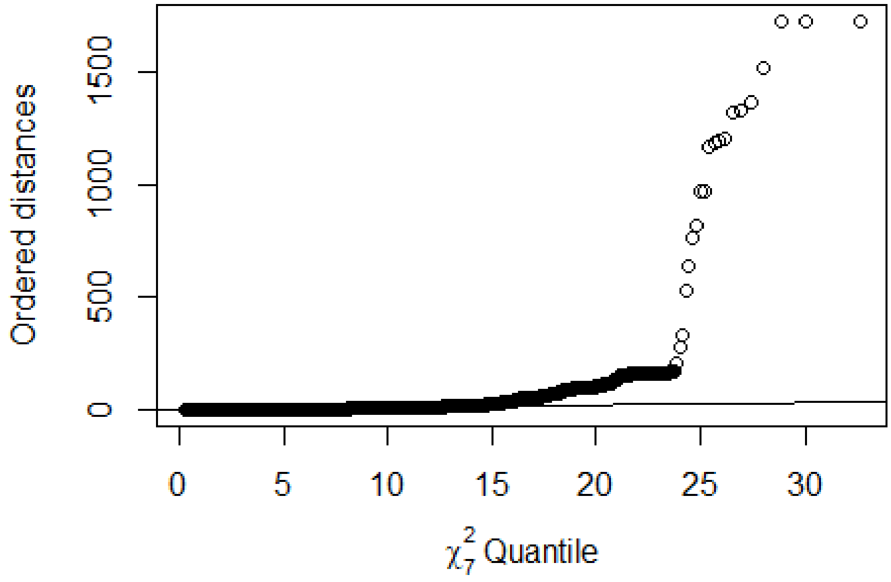
```
q2 = qqnorm(Istanbul.2$number_of_reviews)
```



```
skewness(Istanbul.2$price)
```

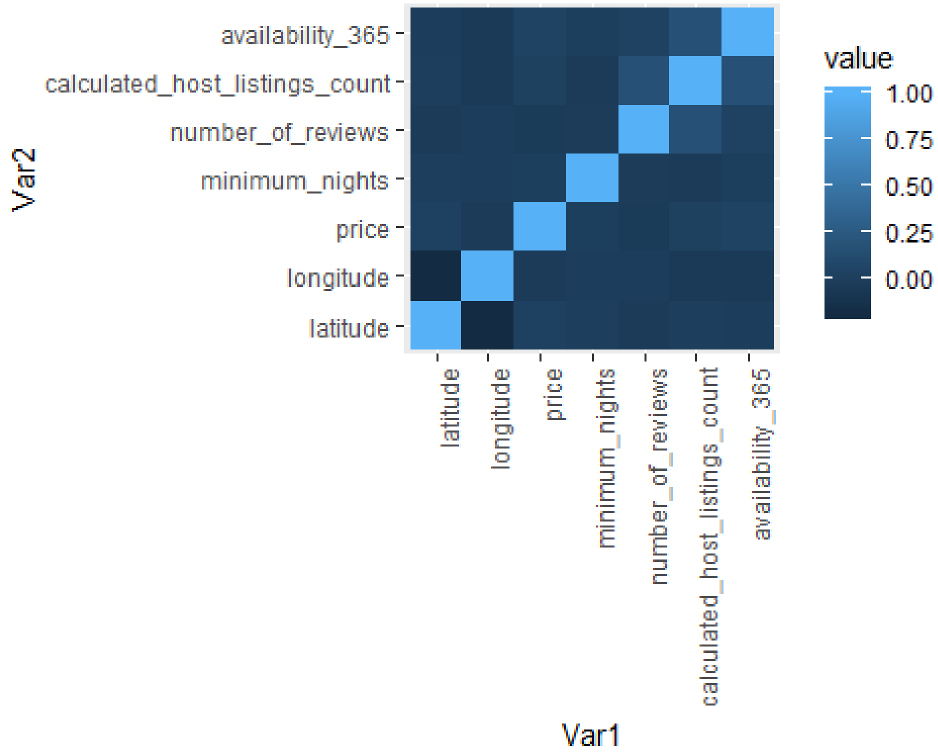
```
## [1] 28.86686
```

```
#skewness is 28.86 so its high skewness as out of range of -1 to 1
#Price not normal

Istanbul.3 <- Istanbul.2[,
c("latitude","longitude","price","minimum_nights","number_of_reviews","calcul
ated_host_listings_count","availability_365")]
Istanbul.cm <- colMeans(Istanbul.3) ## average of all the 7 variables
Istanbul.S <- cov(Istanbul.3) ## covariance of the 7 vars
Istanbul.d <- apply(Istanbul.3, MARGIN = 1, function(Istanbul.3)t(Istanbul.3
- Istanbul.cm) %*% solve(Istanbul.S) %*% (Istanbul.3 - Istanbul.cm))

## multi variate chi square plot ## to signify whether my variables are
normally distributed
plot(qchisq((1:nrow(Istanbul.3) - 1/2) / nrow(Istanbul.3), df = 7),
sort(Istanbul.d),
     xlab = expression(paste(chi[7]^2, " Quantile")),
     ylab = "Ordered distances")
abline(a = 0, b = 1)
```



```
## this signifies that multi variables are not normally distributed

#Correlation matrix
corr <- cor(Istanbul.3)
corrmelt <- melt(corr)
```

```
## Warning in melt(corr): The melt generic in data.table has been passed a
matrix
## and will attempt to redirect to the relevant reshape2 method; please note
that
## reshape2 is deprecated, and this redirection is now deprecated as well. To
## continue using melt methods from reshape2 while both libraries are
attached,
## e.g. melt.list, you can prepend the namespace like reshape2::melt(corr).
In the
## next version, this warning will become an error.

ggplot(cormelt) + geom_tile(aes(Var1, Var2, fill=value)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



```
#Some correlation between Calculated host listing and noof reviews
#Some correlation between Price and calculated_host_listings_count
#Some correlation between Price and availability_365
#Some correlation between calculated_host_listings_count & availability_365
#A bit of relation between Price and Lattitude
```