Group-2

Kumarmangal Roy

5/21/2020

Group-2 Team Members

17220389 Kumarmangal Roy

17006969 Prabavathi Papa Rao

17044032 Kaveenaasvini Chandran

17219523 Ng See Kiat

17220137 Adedigba Stephen Olamilekan

17219152 Ali Alrabeei

Details of Dataset

Title :: Factors that affected the airbnb price in Singapore

Date of Data Collection :: 28 August 2019

Purpose of Dataset :: This data was collected to check the price variation of different Airbnb properties across different regions in Singapore

Dimension Details for DataSet :: 7907x16 (RxC)

Content ::

```
##                            Column                            Description
## 1                              id                          Generic ID Column
## 2                            name                        Generic Property Name
## 3                         host_id                      Generic Host ID Column
## 4                       host_name                           Generic Host Name
## 5             neighbourhood_group                  part/regioin of the city
## 6                   neighbourhood              name of the area of the city
## 7                        latitude                          dictionary meaning
## 8                        longitud                          dictionary meaning
## 9                       room type                    types of room available
## 10                          price           price per room per night in SGD
## 11                minimum_nights     minimum nights property being booked
## 12              number_of_reviews                    total number of reviews
## 13                    last_review                        date of last review
## 14              reviews_per_month           % of reviews received per month
## 15 calculated_host_listings_count unique property ID's registered per host
## 16               availability_365          days property available for rent
```

Structure of Dataset ::

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
## Rows: 7,907
## Columns: 16
## $ id                              <int> 49091, 50646, 56334, 71609, 71896, 7...
## $ name                            <chr> "COZICOMFORT LONG TERM STAY ROOM 2",...
## $ host_id                         <int> 266763, 227796, 266763, 367042, 3670...
## $ host_name                       <chr> "Francesca", "Sujatha", "Francesca",...
## $ neighbourhood_group             <chr> "North Region", "Central Region", "N...
## $ neighbourhood                   <chr> "Woodlands", "Bukit Timah", "Woodlan...
## $ latitude                        <dbl> 1.44255, 1.33235, 1.44246, 1.34541, ...
## $ longitude                       <dbl> 103.7958, 103.7852, 103.7967, 103.95...
## $ room_type                       <chr> "Private room", "Private room", "Pri...
## $ price                           <int> 83, 81, 69, 206, 94, 104, 208, 50, 5...
## $ minimum_nights                  <int> 180, 90, 6, 1, 1, 1, 1, 90, 90, 90, ...
## $ number_of_reviews               <int> 1, 18, 20, 14, 22, 39, 25, 174, 198,...
## $ last_review                     <chr> "2013-10-21", "2014-12-26", "2015-10...
## $ reviews_per_month               <dbl> 0.01, 0.28, 0.20, 0.15, 0.22, 0.38, ...
## $ calculated_host_listings_count <int> 2, 1, 2, 9, 9, 9, 9, 4, 4, 4, 32, 32...
## $ availability_365                <int> 365, 365, 365, 353, 355, 346, 172, 5...
```

Summary for the data

```
##        id                name              host_id           host_name
##  Min.   :   49091   Length:7907        Min.   :    23666   Length:7907
##  1st Qu.:15821800   Class :character   1st Qu.: 23058075   Class :character
##  Median :24706270   Mode  :character   Median : 63448912   Mode  :character
##  Mean   :23388625                      Mean   : 91144807
##  3rd Qu.:32348500                      3rd Qu.:155381142
##  Max.   :38112762                      Max.   :288567551
##
##  neighbourhood_group neighbourhood         latitude       longitude
##  Length:7907         Length:7907        Min.   :1.244   Min.   :103.6
##  Class :character    Class :character   1st Qu.:1.296   1st Qu.:103.8
##  Mode  :character    Mode  :character   Median :1.311   Median :103.8
##                                         Mean   :1.314   Mean   :103.8
##                                         3rd Qu.:1.322   3rd Qu.:103.9
##                                         Max.   :1.455   Max.   :104.0
##
##   room_type             price         minimum_nights    number_of_reviews
##  Length:7907        Min.   :    0.0   Min.   :   1.00   Min.   :  0.00
##  Class :character   1st Qu.:   65.0   1st Qu.:   1.00   1st Qu.:  0.00
##  Mode  :character   Median :  124.0   Median :   3.00   Median :  2.00
##                     Mean   :  169.3   Mean   :  17.51   Mean   : 12.81
##                     3rd Qu.:  199.0   3rd Qu.:  10.00   3rd Qu.: 10.00
##                     Max.   :10000.0   Max.   :1000.00   Max.   :323.00
##
##  last_review        reviews_per_month calculated_host_listings_count
##  Length:7907        Min.   : 0.010    Min.   :  1.00
##  Class :character   1st Qu.: 0.180    1st Qu.:  2.00
##  Mode  :character   Median : 0.550    Median :  9.00
##                     Mean   : 1.044    Mean   : 40.61
##                     3rd Qu.: 1.370    3rd Qu.: 48.00
##                     Max.   :13.000    Max.   :274.00
##                     NA's   :2758
##  availability_365
##  Min.   :  0.0
##  1st Qu.: 54.0
##  Median :260.0
##  Mean   :208.7
##  3rd Qu.:355.0
##  Max.   :365.0
##
```

HMISC - Shows Elaborate description specially used for Categoriacal Variable

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
## SGAIRBNB$room_type
##        n  missing distinct
##     7907        0        3
##
## Value      Entire home/apt    Private room     Shared room
## Frequency             4132            3381             394
## Proportion           0.523           0.428           0.050
```

```
## SGAIRBNB$neighbourhood_group
##        n  missing distinct
##     7907        0        5
##
## lowest : Central Region    East Region       North-East Region North Region      West Region
## highest: Central Region    East Region       North-East Region North Region      West Region
##
## Value         Central Region       East Region North-East Region
## Frequency               6309               508               346
## Proportion             0.798             0.064             0.044
##
## Value           North Region       West Region
## Frequency                204               540
## Proportion             0.026             0.068
```

```
## SGAIRBNB$price
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     7907        0      374        1    169.3    153.6       35       44
##      .25      .50      .75      .90      .95
##       65      124      199      300      381
##
## lowest :     0    14    15    18    19, highest:  6000  6944  7000  8900 10000
```

PSYCH - Show eleborate description for continuous data with added feature of grouping

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:Hmisc':
##
##     describe
```

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

```
##
##  Descriptive statistics by group
## group: Entire home/apt
##    vars    n mean     sd median trimmed   mad min   max range  skew kurtosis
## X1    1 4132  227 330.92    178   190.9 83.03   0 10000 10000 19.16   473.54
##      se
## X1 5.15
## ------------------------------------------------------------
## group: Private room
##    vars    n   mean     sd median trimmed   mad min   max range  skew kurtosis
## X1    1 3381 110.94 353.88     69   76.06 31.13  14 10000  9986 20.38   484.42
##      se
## X1 6.09
## ------------------------------------------------------------
## group: Shared room
##    vars   n  mean     sd median trimmed   mad min  max range  skew kurtosis
## X1    1 394 65.68 157.65     33   38.92 11.86  14 2500  2486 10.91   149.95
##      se
## X1 7.94
```

Checking which data needs to be tidy:

```
# Check missingness
colMeans(is.na(SGAIRBNB))
```

```
##                             id                           name
##                      0.0000000                      0.0000000
##                        host_id                      host_name
##                      0.0000000                      0.0000000
##            neighbourhood_group                  neighbourhood
##                      0.0000000                      0.0000000
##                       latitude                      longitude
##                      0.0000000                      0.0000000
##                      room_type                          price
##                      0.0000000                      0.0000000
##                minimum_nights              number_of_reviews
##                      0.0000000                      0.0000000
##                    last_review             reviews_per_month
##                      0.0000000                      0.3488049
## calculated_host_listings_count               availability_365
##                      0.0000000                      0.0000000
```

```
sum(is.na(SGAIRBNB$reviews_per_month))
```

```
## [1] 2758
```

```
#We have got 2758 missing values its not logical to impute so many values so we are dropping the parameter. So its logical to remove the particular column


#Checking basic data
unique(SGAIRBNB$neighbourhood) #there are 43 specific area in the city
```

```
##  [1] "Woodlands"               "Bukit Timah"
##  [3] "Tampines"                "Bedok"
##  [5] "Bukit Merah"             "Newton"
##  [7] "Geylang"                 "River Valley"
##  [9] "Jurong West"             "Rochor"
## [11] "Queenstown"              "Serangoon"
## [13] "Marine Parade"           "Pasir Ris"
## [15] "Toa Payoh"               "Outram"
## [17] "Punggol"                 "Tanglin"
## [19] "Hougang"                 "Kallang"
## [21] "Novena"                  "Downtown Core"
## [23] "Bukit Panjang"           "Singapore River"
## [25] "Orchard"                 "Ang Mo Kio"
## [27] "Bukit Batok"             "Museum"
## [29] "Sembawang"               "Choa Chu Kang"
## [31] "Central Water Catchment" "Sengkang"
## [33] "Clementi"                "Jurong East"
## [35] "Bishan"                  "Yishun"
## [37] "Mandai"                  "Southern Islands"
## [39] "Sungei Kadut"            "Western Water Catchment"
## [41] "Tuas"                    "Marina South"
## [43] "Lim Chu Kang"
```

```
unique(SGAIRBNB$neighbourhood_group) #mapped to 5 broader regions
```

```
## [1] "North Region"      "Central Region"    "East Region"
## [4] "West Region"       "North-East Region"
```

```
unique(SGAIRBNB$room_type) #3 type of rental available
```

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

```
summary(SGAIRBNB$price) #0 price is not possible for any property
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    65.0   124.0   169.3   199.0 10000.0
```

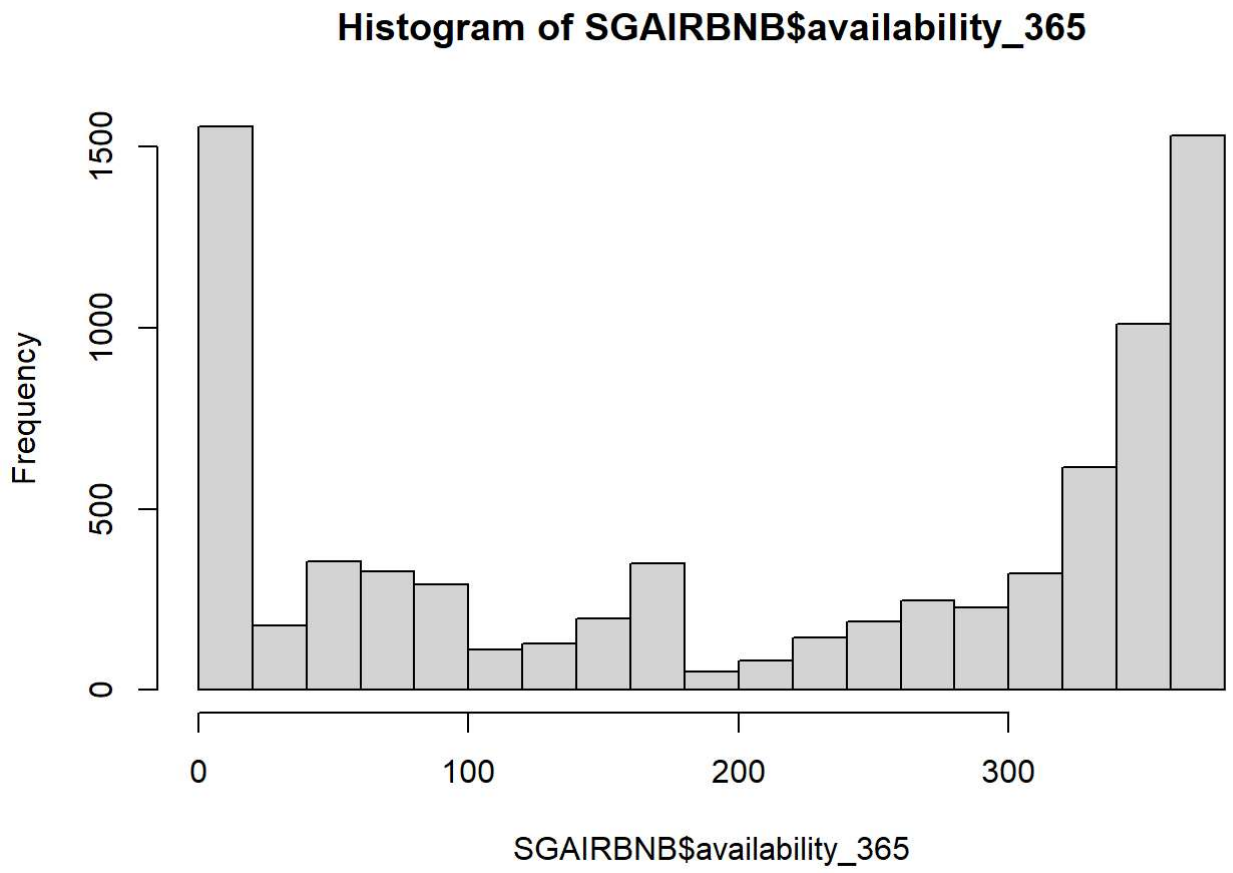
```
sum(SGAIRBNB$price == 0) #Need to remove 1 data point
```

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

```
SGAIRBNB <- subset(SGAIRBNB,SGAIRBNB$price != 0)

#There are some wrong data type that we need to change it into corect type, and save it into new
object 'airbnb'.Converting ids into factor

# Checking the distribution of a column
# No apparent problem in the distribution of season, noting a downward trend in records as time
 increases though
hist(SGAIRBNB$availability_365)
```




```
# Checking if one row is identical to another
distinctdata <- distinct(SGAIRBNB)
nrow(SGAIRBNB)
```

```
## [1] 7906
```

```
# The dataset of distinct values has the same number of rows as the original, meaning there are
 no duplicates!
nrow(distinctdata)
```

```
## [1] 7906
```

Preparing data for analysis by correcting the variables and contents of the data.:

```
SGAIRBNB <- select(SGAIRBNB, -reviews_per_month) #Remove the particular column with missing vari
able

SGAIRBNB <- subset(SGAIRBNB,SGAIRBNB$price != 0) #Removing only 1 data point

#There are some wrong data type that we need to change it into corect type, and save it into new
object 'airbnb'.Converting ids into factor

SGAIRBNB$id <- as.factor(SGAIRBNB$id)
SGAIRBNB$host_id <- as.factor(SGAIRBNB$host_id)
SGAIRBNB$neighbourhood_group <- as.factor(SGAIRBNB$neighbourhood_group)
SGAIRBNB$neighbourhood <- as.factor(SGAIRBNB$neighbourhood)
SGAIRBNB$room_type <- as.factor(SGAIRBNB$room_type)
str(SGAIRBNB)
```

```
## 'data.frame':    7906 obs. of  15 variables:
##  $ id                            : Factor w/ 7906 levels "49091","50646",..: 1 2 3 4 5 6 7 8
9 10 ...
##  $ name                          : chr  "COZICOMFORT LONG TERM STAY ROOM 2" "Pleasant Room al
ong Bukit Timah" "COZICOMFORT" "Ensuite Room (Room 1 & 2) near EXPO" ...
##  $ host_id                       : Factor w/ 2705 levels "23666","59498",..: 10 5 10 15 15 15
15 39 39 39 ...
##  $ host_name                     : chr  "Francesca" "Sujatha" "Francesca" "Belinda" ...
##  $ neighbourhood_group           : Factor w/ 5 levels "Central Region",..: 4 1 4 2 2 2 2 2 2
2 ...
##  $ neighbourhood                 : Factor w/ 43 levels "Ang Mo Kio","Bedok",..: 42 7 42 37 37
37 37 2 2 2 ...
##  $ latitude                      : num  1.44 1.33 1.44 1.35 1.35 ...
##  $ longitude                     : num  104 104 104 104 104 ...
##  $ room_type                     : Factor w/ 3 levels "Entire home/apt",..: 2 2 2 2 2 2 2 2 2
2 ...
##  $ price                         : int  83 81 69 206 94 104 208 50 54 42 ...
##  $ minimum_nights                : int  180 90 6 1 1 1 1 90 90 90 ...
##  $ number_of_reviews             : int  1 18 20 14 22 39 25 174 198 236 ...
##  $ last_review                   : chr  "2013-10-21" "2014-12-26" "2015-10-01" "2019-08-11"
...
##  $ calculated_host_listings_count: int  2 1 2 9 9 9 9 4 4 4 ...
##  $ availability_365              : int  365 365 365 353 355 346 172 59 133 147 ...
```

Objective/goal of processing this dataset ::

1. Price Factor - Question : Which area is the most populated for high rental cost? – here we can add in the reason why it is expensive, maybe it is close to Marina Sands bay or shopping malls.
2. Listing type Factor – Question : What is the most demanding type for staying in Singapore? -Here we can have range of price (like how we discussed in previous call) and then identify what is s the most demanding type of listing (like private room/apt/shared room) on Singapore Airbnb.
3. Popular Area Factor– Question : Question: Where is the most demanding area in Central Region ?-here to know which is the popular are in that region.

Cleaned Processed Dataset ::

```
##      id                              name host_id host_name neighbourhood_group
## 1 49091 COZICOMFORT LONG TERM STAY ROOM 2  266763 Francesca        North Region
## 2 50646   Pleasant Room along Bukit Timah  227796   Sujatha      Central Region
## 3 56334                      COZICOMFORT  266763 Francesca        North Region
##   neighbourhood latitude longitude    room_type price minimum_nights
## 1     Woodlands  1.44255  103.7958 Private room    83            180
## 2   Bukit Timah  1.33235  103.7852 Private room    81             90
## 3     Woodlands  1.44246  103.7967 Private room    69              6
##   number_of_reviews last_review calculated_host_listings_count availability_365
## 1                 1  2013-10-21                              2              365
## 2                18  2014-12-26                              1              365
## 3                20  2015-10-01                              2              365
```