Singapore Airbnb Analysis

Chan Wei Wei 16052748 , Chan Wei Chee 16052755

R Markdown

The dataset is derived from **Inside Airbnb** which consists of two files:

- Singapore Airbnb Listings October2020
- Singapore Airbnb Reviews October 2020

A.Introduction and motivation of the work

Airbnb is an online marketplace which allows local hosts to rent their properties or spare rooms to the guests.Airbnb currently covers more than 100,000 cities and 220 countries worldwide.For this assignment, we will be looking into the **Airbnb listings in Singapore as of year 2020**.

Singapore is undoubtedly one of the best cities for short-term travelling in Southeast Asia with most efficient public transport system,affordable dining and environment. Staying in an Airbnb property has always been the go-to option for many travellers since Airbnb offers a comparable pricing to many of Singapore's best hostels.

Our team have chosen to work on the Singapore Airbnb dataset since Singapore is our neighbour country and having similar living culture to Malaysia.Likewise,we can better relate and understand our analysis.Also, we are motivated to deal with the dataset regarding Airbnb as we are interested to start hosting Airbnb in the future.Hence,the information provided about the Airbnb hosts,listings and guests are worth exploring.

Standpoint from a traveller,this analysis will provide an overview of the geographic and demographic inforamation about the Airbnb listings in Singapore.Questions like **which neighbourhood contains the most Airbnb lisings** and **which neighbourhood is the most expensive and cheapest to book on Airbnb** can be answered through the analysis carried out in the following sections.

Besides,the Sinagpore Airbnb dataset picks up our curiosity to dig further down about the market price and market demand around each region of Singapore.For example,we would like to know about **the average price by room type and neighbourhood group in Singapore.**

Moreover,this analysis could be helpful for the hosts to better understand their guest's expectation through **text mining and sentiment analysis on the reviews.**

```
#set working directory
setwd("C:/Users/user/Documents/R project/Singapore Airbnb/IST2334-
16052748&16052755")
```

```
library(cowplot)

## Warning: package 'cowplot' was built under R version 4.0.3

library(ggplot2)
library(dplyr)

##
## 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

library(glue)

##
## Attaching package: 'glue'

## The following object is masked from 'package:dplyr':
##
##     collapse
```

B.Elaboration of the dataset

The Singapore Airbnb dataset consists of 4492 rows and 16 variables which depicts the host listing information.However,only 8 variables that related to our analysis have been selected.The summary table shows that there is no missing value presented in these variables. Below are the selected variables with their respective descriptions:

- **name**- Title of Listings
- **neighbourhood_group**- Region that contains listing
- **neighbourhood**- Name of neighbourhood
- **latitude**- The geographical information
- **longitude**- The geographical information
- **room_type**- Type of room(Entire home/apt,Private room,Shared room,Hotel room)
- **price**- Price per night (SGD)
- **minimum_nights**- The minimum number of nights required to book the listings

```
#loading the dataset
airbnb <- read.csv("C:/Users/user/Documents/R project/Singapore
Airbnb/listings.csv")

dim(airbnb)

## [1] 4492   16
```

```r
#The airbnb data contains 4492 rows and 16 column

#select only variables related to our analysis
airbnb <- select(airbnb, name,neighbourhood_group, neighbourhood, latitude,
longitude, room_type, price,minimum_nights)

#Basic data exploration
#Viewing the first 10 dataframe records
head(airbnb, 10)
```

```
##                                                 name neighbourhood_group
## 1                 COZICOMFORT LONG TERM STAY ROOM 2        North Region
## 2                   Pleasant Room along Bukit Timah      Central Region
## 3                                        COZICOMFORT        North Region
## 4                Ensuite Room (Room 1 & 2) near EXPO         East Region
## 5                    B&B  Room 1 near Airport & EXPO         East Region
## 6                         Room 2-near Airport & EXPO         East Region
## 7                   3rd level Jumbo room 5 near EXPO         East Region
## 8             Long stay at The Breezy East "Leopard"         East Region
## 9            Long stay at The Breezy East "Plumeria"         East Region
## 10 Conveniently located City Room!(1,2,3,4,5,6,7,8)      Central Region
##    neighbourhood latitude longitude    room_type price minimum_nights
## 1      Woodlands  1.44255  103.7958 Private room    82            180
## 2    Bukit Timah  1.33235  103.7852 Private room    80             90
## 3      Woodlands  1.44246  103.7967 Private room    68              6
## 4       Tampines  1.34541  103.9571 Private room   179             90
## 5       Tampines  1.34567  103.9596 Private room    95             90
## 6       Tampines  1.34702  103.9610 Private room    82             90
## 7       Tampines  1.34348  103.9634 Private room   208              1
## 8          Bedok  1.32391  103.9128 Private room    52             90
## 9          Bedok  1.32391  103.9128 Private room    54             90
## 10   Bukit Merah  1.28875  103.8081 Private room    52             14
```

```r
summary(airbnb)
```

```
##      name           neighbourhood_group neighbourhood         latitude
##  Length:4492        Length:4492         Length:4492        Min.   :1.245
##  Class :character   Class :character    Class :character   1st Qu.:1.295
##  Mode  :character   Mode  :character    Mode  :character   Median :1.310
##                                                            Mean   :1.313
##                                                            3rd Qu.:1.322
##                                                            Max.   :1.453
##    longitude      room_type             price         minimum_nights
##  Min.   :103.6   Length:4492        Min.   :   14.0   Min.   :   1.00
##  1st Qu.:103.8   Class :character   1st Qu.:   61.0   1st Qu.:   2.00
##  Median :103.9   Mode  :character   Median :  113.0   Median :   6.00
##  Mean   :103.8                      Mean   :  163.2   Mean   :  26.19
##  3rd Qu.:103.9                      3rd Qu.:  170.0   3rd Qu.:  28.00
##  Max.   :104.0                      Max.   :10286.0   Max.   :1000.00
```

```r
c(unique(airbnb["neighbourhood_group"]))
```

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

```
c(unique(airbnb["neighbourhood"]))
```

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

```
c(unique(airbnb["room_type"]))
```

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

```
glue ("Price Minumum:{min(airbnb$price)} |Price Maximum:{max(airbnb$price)}")
```

```
## Price Minumum:14 |Price Maximum:10286
```

```
glue ("Night Minumum:{min(airbnb$minimum_nights)} |Night
Maximum:{max(airbnb$minimum_nights)}")
```

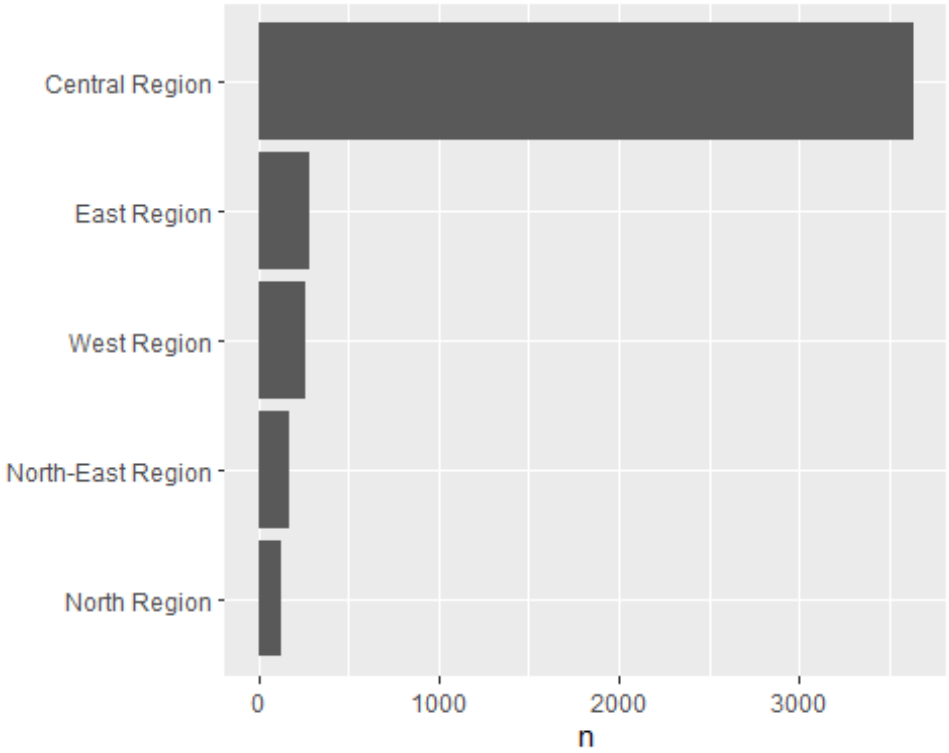
```
## Night Minumum:1 |Night Maximum:1000
```

From the preliminary data exploration, we get to understand that the airbnb listings are distributed into **5 neighbourhood_group** namely, ***North***,***Central***,***East***,***North-East*** and ***West*** Region. Within these regions,there are a total of **44 neighbourhoods**.Also,there are four type of room options available which are ***private room***,***entire home or apartment***,***shared room*** and ***hotel room***.The price ranges from **as low as SGD 14 and to as high as SGD 10286.**

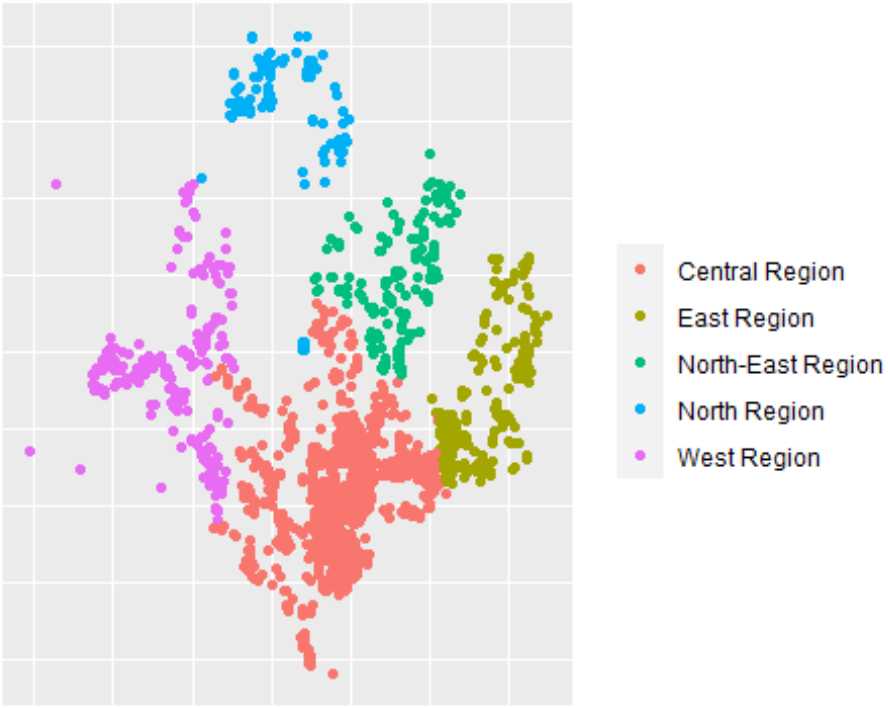
Frequency Distribution

i.Count of neighbourhood group per region

```
#construct frequency table of neighbourhood group
freq_location <- data.frame(cbind(Frequency =
table(airbnb$neighbourhood_group), Percent =
prop.table(table(airbnb$neighbourhood_group)) * 100))
freq_location <- freq_location[order(freq_location$Frequency),]
freq_location

##                    Frequency   Percent
## North Region             129  2.871772
## North-East Region        173  3.851291
## West Region              263  5.854853
## East Region              289  6.433660
## Central Region          3638 80.988424

#frequency plot of neighbourhood group
airbnb %>%
  count(neighbourhood_group, sort = TRUE) %>%
  filter(n > 100) %>%
  mutate(neighbourhood_group = reorder(neighbourhood_group, n)) %>%
  ggplot(aes(neighbourhood_group, n)) +
    geom_col() +
    xlab(NULL) +
    coord_flip()
```



```r
#geographical map of neighbourhood_group
ggplot(data = airbnb) +
  geom_point(mapping = aes(x = longitude, y = latitude,
color=neighbourhood_group)) +
  xlab("") +
  ylab("") +
  labs(color = NULL) +
  theme(axis.text.x = element_blank(), axis.text.y = element_blank(),
axis.ticks = element_blank())
```



In this section,we would like to understand the frequency distribution of the neighbourhood group.The geographical map clearly shows that majority of the listings(81%) are densely populated at the Central Region,followed by East region(6.4%),West Region(5.9%),North-East Region(3.9%) and North Region(2.9%).

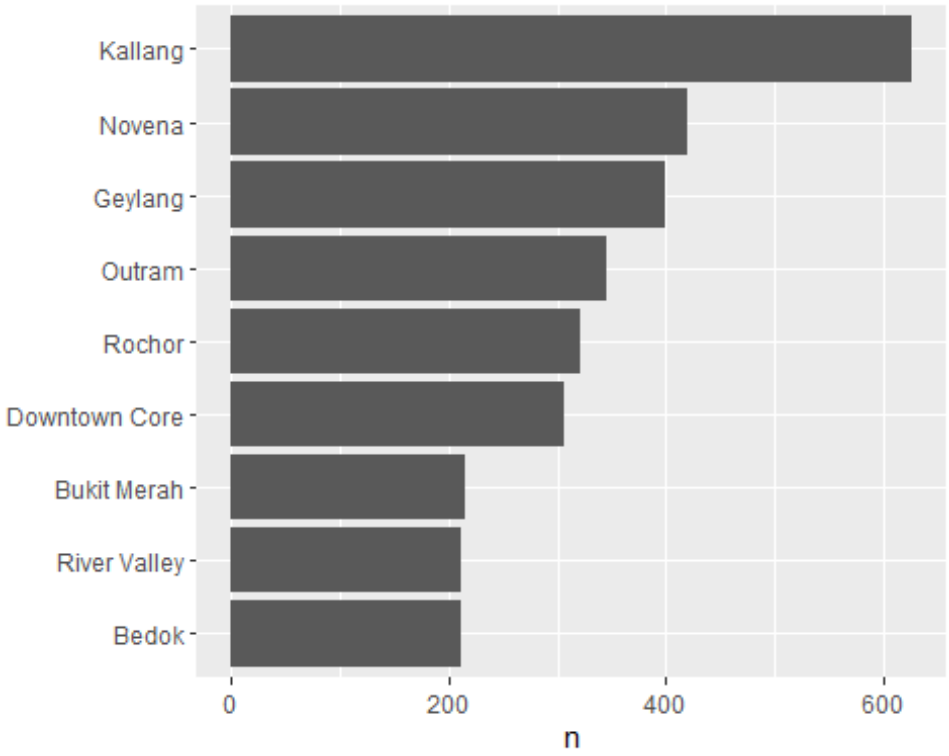
There is no doubt that Central region occupied most of the Airbnb listings as the main city is strategically located around the major tourist attractions.For example,the landmark of Singapore - Merlion Park,Marina Bay Sands,Gardens by the Bay are located at Downtown Core District.

ii.Count of neighbourhood per area

```r
#construct frequency table of neighbourhood
freq_area <- data.frame(cbind(Frequency = table(airbnb$neighbourhood),
Percent = prop.table(table(airbnb$neighbourhood)) * 100))
freq_area <- freq_area[order(freq_area$Frequency),]
freq_area
```

```
##                            Frequency    Percent
## Pioneer                            1  0.0222618
## Sungei Kadut                       1  0.0222618
## Tuas                               1  0.0222618
## Changi                             2  0.0445236
## Mandai                             2  0.0445236
## Marina South                       2  0.0445236
## Western Water Catchment            2  0.0445236
## Central Water Catchment           12  0.2671416
## Punggol                           18  0.4007124
## Choa Chu Kang                     21  0.4674978
## Bukit Panjang                     22  0.4897596
## Sengkang                          23  0.5120214
## Yishun                            24  0.5342832
## Ang Mo Kio                        25  0.5565450
## Southern Islands                  26  0.5788068
## Museum                            29  0.6455922
## Pasir Ris                         29  0.6455922
## Bukit Batok                       36  0.8014248
## Sembawang                         38  0.8459484
## Bishan                            39  0.8682102
## Serangoon                         42  0.9349955
## Tampines                          47  1.0463045
## Jurong East                       52  1.1576135
## Woodlands                         52  1.1576135
## Marine Parade                     54  1.2021371
## Jurong West                       59  1.3134461
## Hougang                           65  1.4470169
## Newton                            65  1.4470169
## Clementi                          69  1.5360641
## Bukit Timah                       74  1.6473731
## Toa Payoh                         74  1.6473731
## Orchard                           78  1.7364203
## Tanglin                           85  1.8922529
## Queenstown                       120  2.6714159
## Singapore River                  147  3.2724844
## Bedok                            211  4.6972395
## River Valley                     211  4.6972395
## Bukit Merah                      215  4.7862867
## Downtown Core                    306  6.8121104
## Rochor                           321  7.1460374
## Outram                           346  7.7025824
## Geylang                          400  8.9047195
## Novena                           419  9.3276937
## Kallang                          627 13.9581478

#frequency plot of neighbourhood
airbnb %>%
  count(neighbourhood, sort = TRUE) %>%
  filter(n > 150) %>%
```

```
  mutate(neighbourhood = reorder(neighbourhood, n)) %>%
  ggplot(aes(neighbourhood, n)) +
    geom_col() +
    xlab(NULL) +
    coord_flip()
```

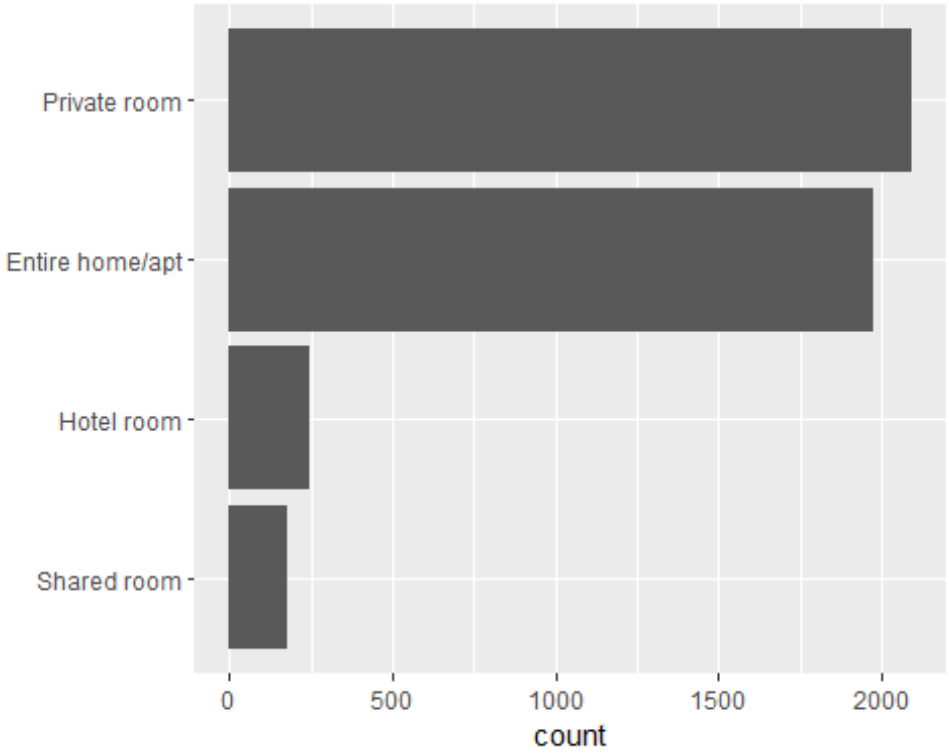



```
#geographical map of neighbourhood
ggplot(data = airbnb) +
  geom_point(mapping = aes(x = longitude, y = latitude, color=neighbourhood))
+
  xlab("") +
  ylab("") +
  labs(color = NULL) +
  theme(axis.text.x = element_blank(), axis.text.y = element_blank(),
axis.ticks = element_blank())
```

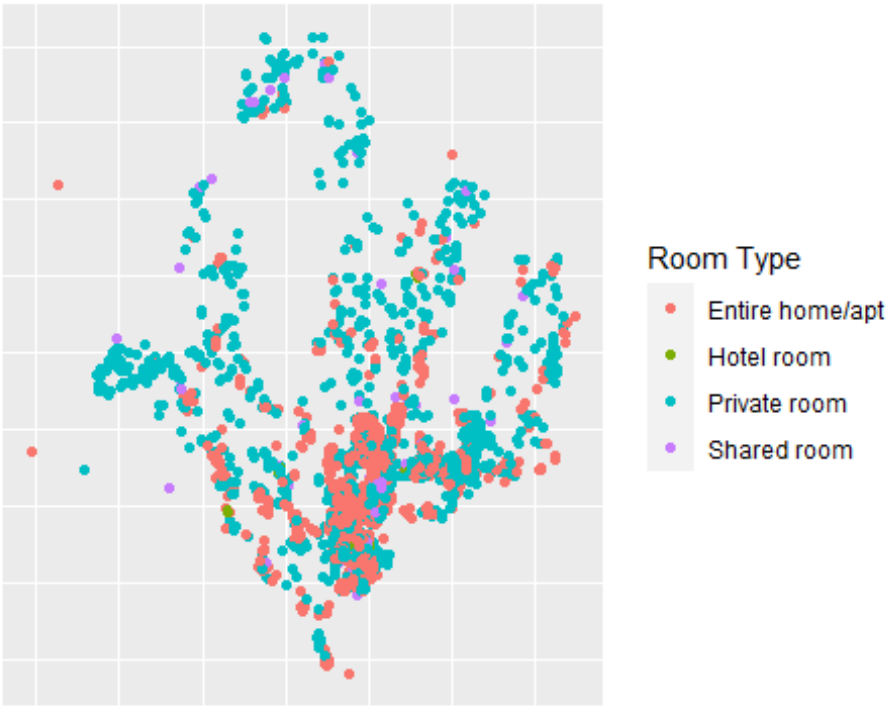
The Airbnb listings in the central region are more concentrated at these three neighbourhood : Kallang,Novena and Geylang.

iii.Count of Room type

```
#construct frequency table of room type
freq_type <- data.frame(cbind(Frequency = table(airbnb$room_type), Percent =
prop.table(table(airbnb$room_type)) * 100))
freq_type <- freq_type[order(freq_type$Frequency),]
freq_type

##                 Frequency   Percent
## Shared room           176  3.918077
## Hotel room            248  5.520926
## Entire home/apt      1974 43.944791
## Private room         2094 46.616207

#frequency plot of room type
room_type <-  airbnb %>%
  count(room_type, sort = TRUE) %>%
  mutate(room_type = reorder(room_type, n)) %>%
  ggplot(aes(room_type, n)) +
    geom_col() +
    xlab(NULL) +
    ylab("count") +
    coord_flip()

plot_grid(room_type, nrow = 1)
```



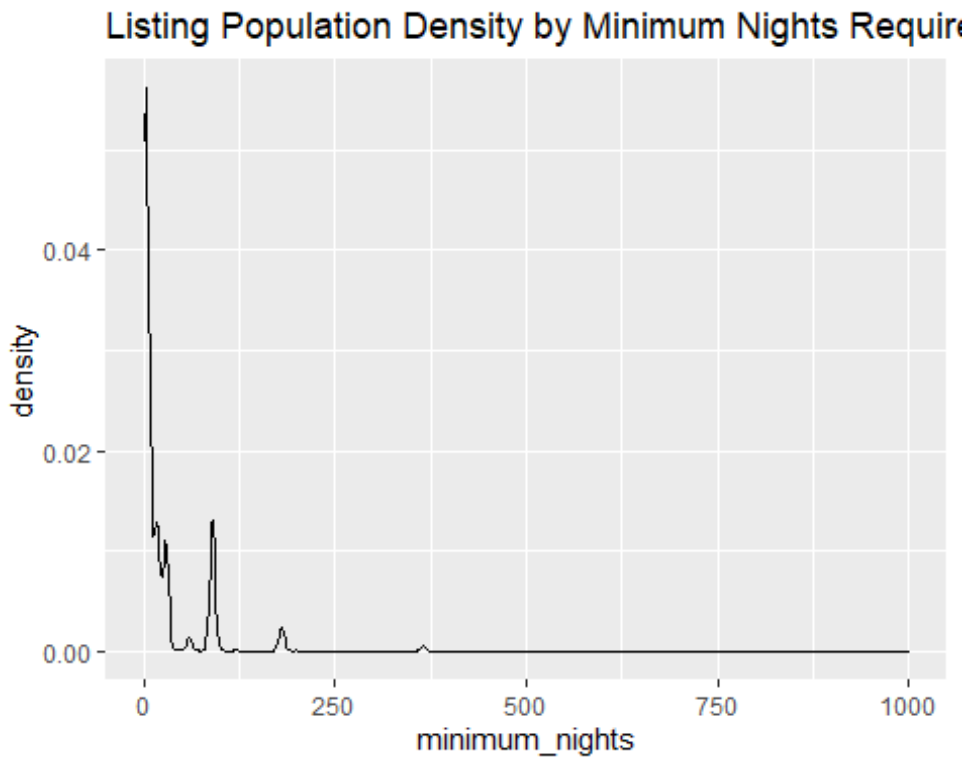

```
#geographical map of room type by location
ggplot(data = airbnb) +
  geom_point(mapping = aes(x = longitude, y = latitude, color=room_type)) +
  xlab("") +
  ylab("") +
  labs(color = 'Room Type') +
  theme(axis.text.x = element_blank(), axis.text.y = element_blank(),
axis.ticks = element_blank())
```



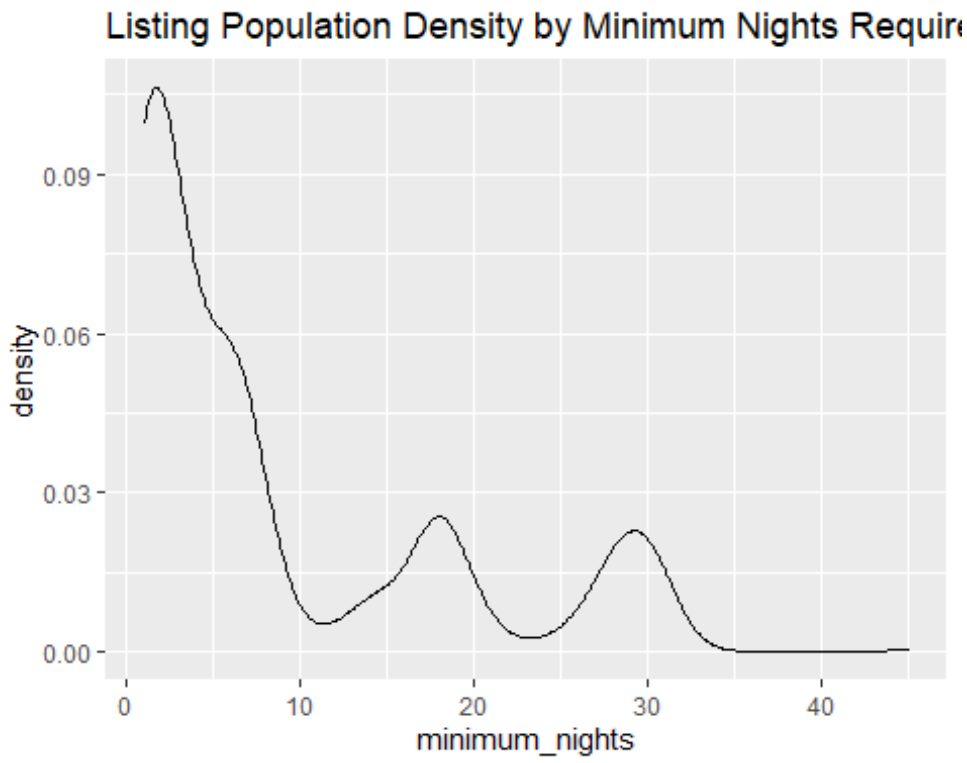
The main room type in the Airbnb listings are private room(46%) and entire home/apartment(43%). The hotel room and shared room contributes below 10 % of the available listings.From the geographical map, it is observed that the home/apartment are located more concentrated at the Central Region whereas private room are disributed around each region.

Distribution of minimum_nights

```
#graph distribution of minimum_nights
ggplot(airbnb, aes(minimum_nights)) + geom_density() + ggtitle("Listing
Population Density by Minimum Nights Required")
```



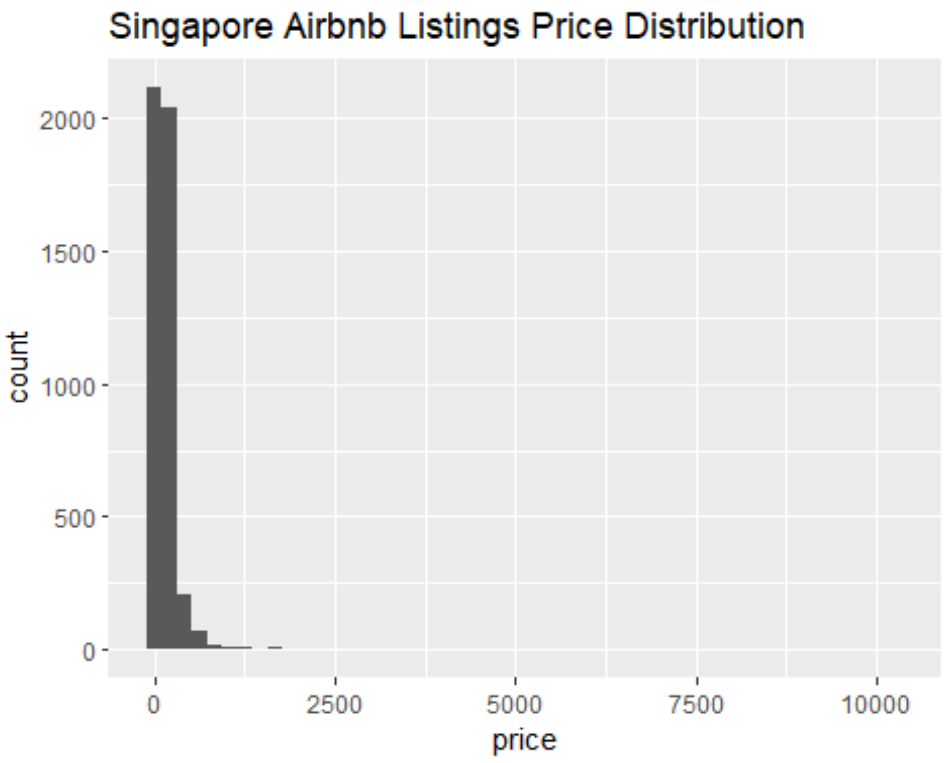

```
ggplot(subset(airbnb, minimum_nights < 50), aes(minimum_nights)) +
geom_density() + ggtitle("Listing Population Density by Minimum Nights
Required")
```



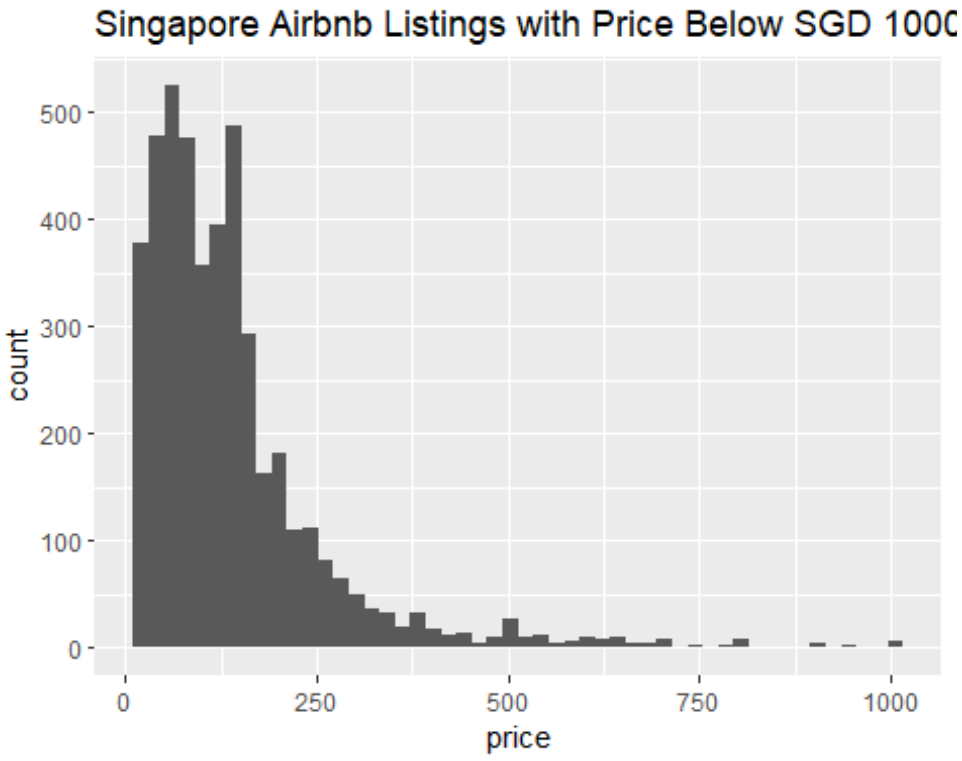
The distribution graph shows that majority of the listings are setting the minimum nights required under 10 nights

Distribution of price

```
#graph distribution of price
price<- airbnb %>%
  ggplot(aes(price)) + geom_histogram(bins=50)+ggtitle("Singapore Airbnb
Listings Price Distribution")
plot_grid(price, nrow = 1)
```




```
#closer look at price below 1000
price<- airbnb %>% filter(price<=1000) %>%
  ggplot(aes(price)) + geom_histogram(bins=50)+ggtitle("Singapore Airbnb
Listings with Price Below SGD 1000")
plot_grid(price, nrow = 1)
```

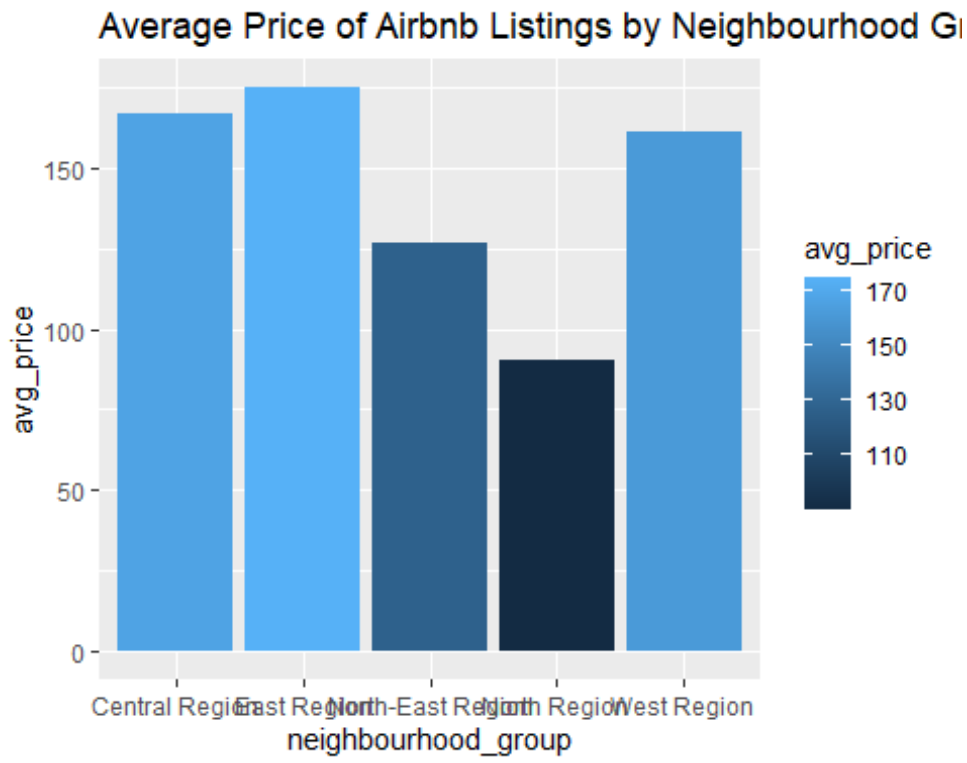


The price distribution of Singapore Airbnb listings are heavily right skewed.The first distribution graph shows that the price of the Airbnb listings are below SGD 2500 per night. However, we narrowed down our scope to the listings with price below SGD 1000 per night.Now, the distribution graph clearly shows that most of the listings are cost below SGD 250 per night.We will look more in depth about the pricing at the next section.

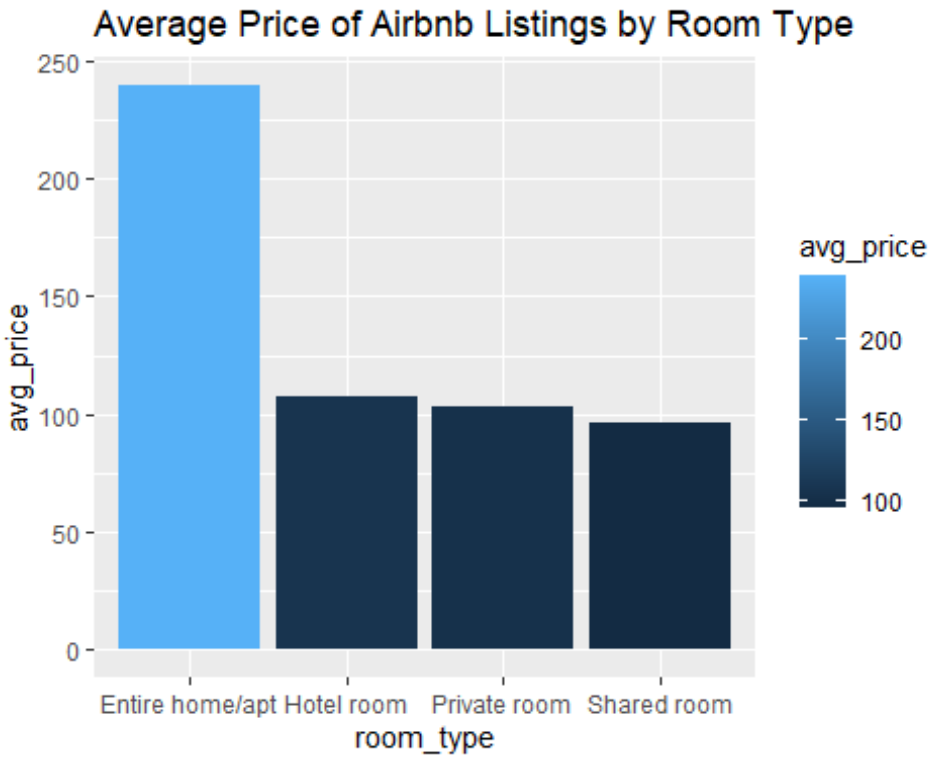
Analysis Question 1:

What is the Average Price of the Airbnb Listings by Room type and Neighbourhood group?

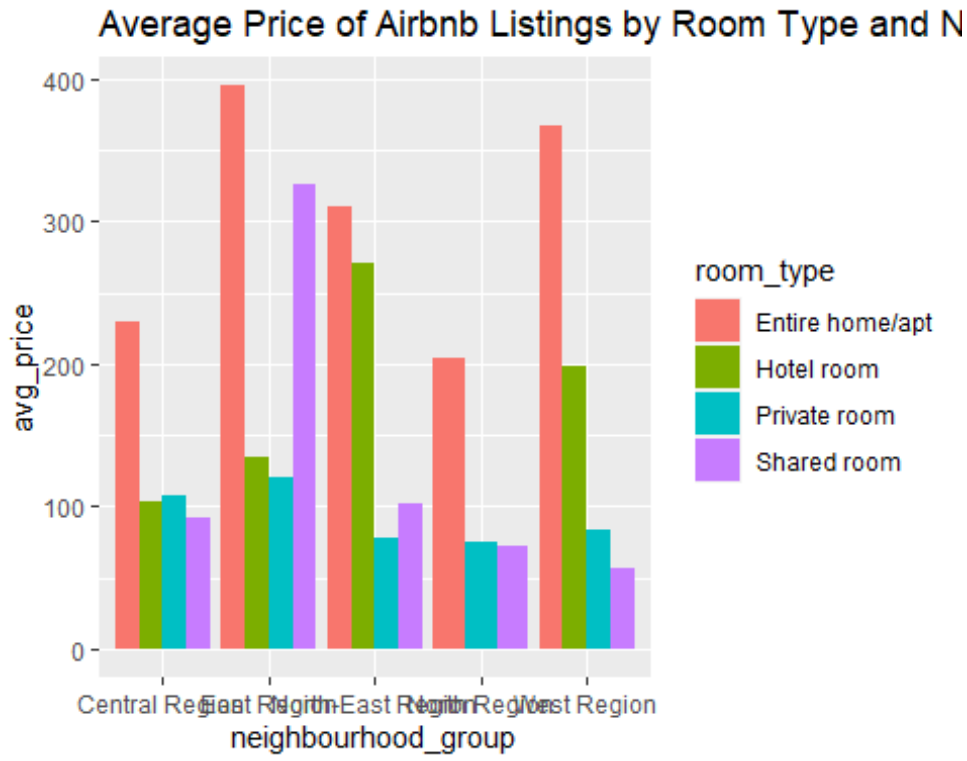
```
#Average price of neighbourhood group
avg_price_group <- airbnb %>%
  group_by(neighbourhood_group) %>%
  summarise(avg_price= mean(price))

## `summarise()` ungrouping output (override with `.groups` argument)

ggplot(avg_price_group , aes(x=neighbourhood_group, y=avg_price)) +
  geom_col(aes(fill=avg_price)) +ggtitle("Average Price of Airbnb Listings by
Neighbourhood Group")
```




```
#Average price of room type
avg_room_price <- airbnb %>%
  group_by(room_type) %>%
  summarise(avg_price= mean(price))

## `summarise()` ungrouping output (override with `.groups` argument)

ggplot(avg_room_price , aes(x=room_type, y=avg_price)) +
  geom_col(aes(fill=avg_price))+ggtitle("Average Price of Airbnb Listings by
Room Type ")
```




```
#Average price of listings by room type and neighbourhood group
avgprice_roomtype_by_region <- airbnb %>%
  group_by(neighbourhood_group,room_type) %>%
  summarise(avg_price= mean(price))

## `summarise()` regrouping output by 'neighbourhood_group' (override with
`.groups` argument)

ggplot(avgprice_roomtype_by_region, aes(x = neighbourhood_group, y =
avg_price, fill = room_type)) +
  geom_col(position = 'dodge')+ggtitle("Average Price of Airbnb Listings by
Room Type and Neighbourhood Group ")
```



The graph shows that the average price of listings based on room type and neighbourhood group.It is observed that the entire home/apartment in East region has the highest average price which costs nearly SGD400 while North Region has the lowest average price which costs around SGD210.

Next,the average price of hotel room costs the most expensive(around SGD280) at North-East Region while having most cheapest price at the Central Region (SGD100).Morevoer,the average price of the private room has not much difference across each region with costs ranging between SGD80 to SGD120.It is interesting to observe that shared room have the highest contrast in its average price.Renting a shared room at the East Region costs around SGD330 which is far more expensive than renting a private room.

Surprisingly, although the Central Region has the most numbers of listings with different room type,its average price are considerably lower compared to the other region.Our **assumption** would be the price of listings are highly competitive at the Central Region.As a result,the airbnb listings could be hard to rent out if the Airbnb owner priced it too high.

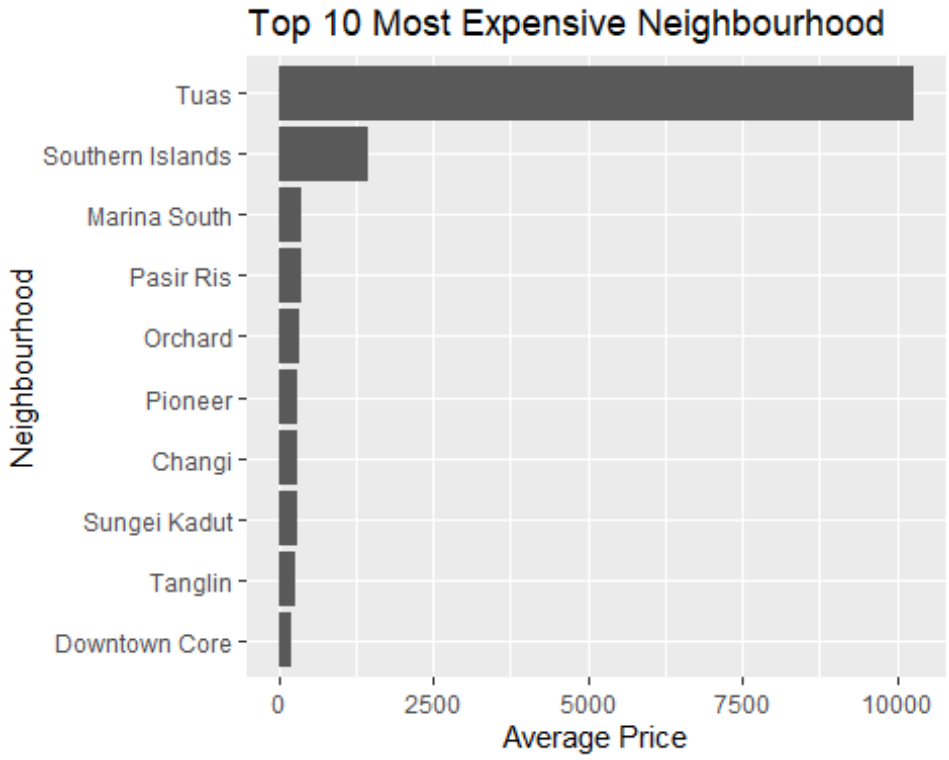
On the other hand,the Airbnb listings at East region and North East Region has considerably high average price. According to wikipedia, the East Region is the second most densely populated among the five and has the smallest land area.Thus,the low supplies but high demands could be the reason leads to high pricing of property in these areas.Therefore, the Airbnb owners have to charge their rentals at a higher rate to cover their expenses.

Analysis Question 2:

Top 10 Most Expensive and Cheapest Neighbourhoods to Book on Airbnb

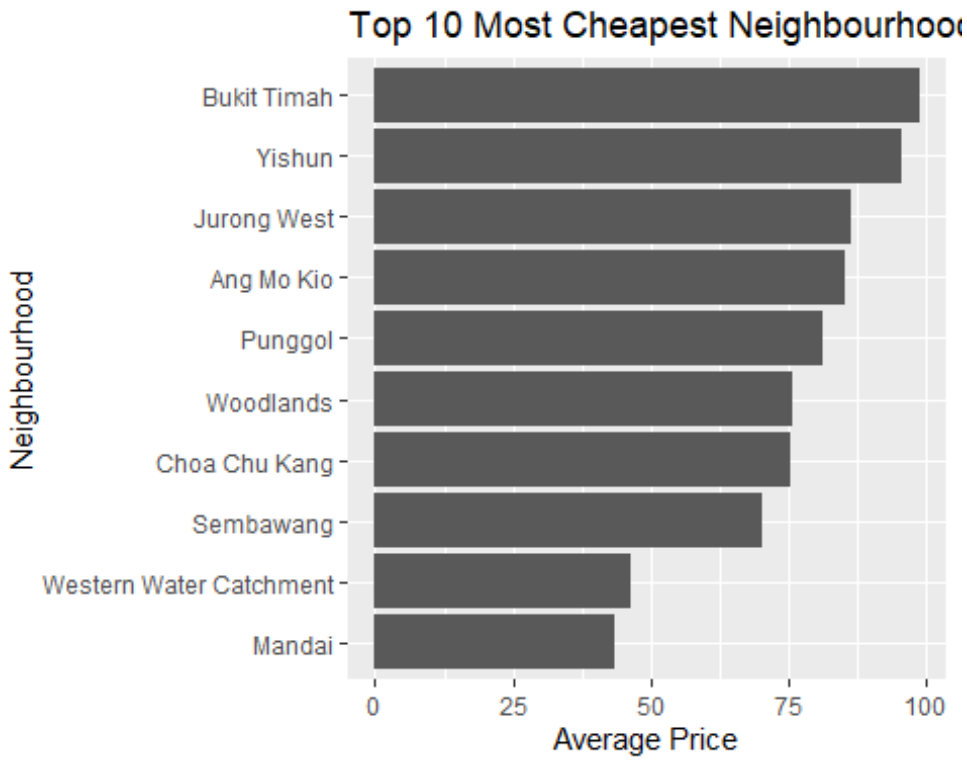
To be more concise,we now zoom in to find out the most expensive and cheapest neighbourhoods to book a room.

```
#Top 10 most expensive neighbourhood
avg_price_exp <- airbnb %>%
  group_by(neighbourhood) %>%
  summarise(avg_price= mean(price)) %>%
  arrange(desc(avg_price))

## `summarise()` ungrouping output (override with `.groups` argument)

ggplot(avg_price_exp[1:10,],aes(x=reorder(neighbourhood,avg_price),
y=avg_price)) +
  geom_bar(stat='identity') + ggtitle('Top 10 Most Expensive Neighbourhood')+
  xlab("Neighbourhood") +
  ylab("Average Price") +
  coord_flip()
```




```
#Top 10 most cheapest neighbourhood
avg_price_cheap <- airbnb %>%
  group_by(neighbourhood) %>%
  summarise(avg_price= mean(price)) %>%
  arrange(avg_price)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

ggplot(avg_price_cheap[1:10,],aes(x=reorder(neighbourhood,avg_price),
y=avg_price)) +
  geom_bar(stat='identity') + ggtitle('Top 10 Most Cheapest Neighbourhood')+
  xlab("Neighbourhood") +
  ylab("Average Price") +
  coord_flip()
```



The list of **Top 10 Most Expensive Neighbourhoods** are as follow:

Central Region:

- **Tuas**
- **Southern Islands**
- **Marina South**
- **Orchard**
- **Pioneer**
- **Tanglin**
- **Downtown Core**

East Region:

- **Pasir Ris**
- **Changi**

North Region:

- **Sungei Kadut**

The list of **Top 10 Most Cheapest Neighbourhoods** are as follow:

North Region:

- **Mandai**
- **Sembawang**
- **Woodlands**
- **Yishun**

West Region:

- **Western Water Catchment**
- **Choa Chu Kang**
- **Jurong West**

North-East Region:

- **Punggol**
- **Ang Mo Kio**

Central Region:

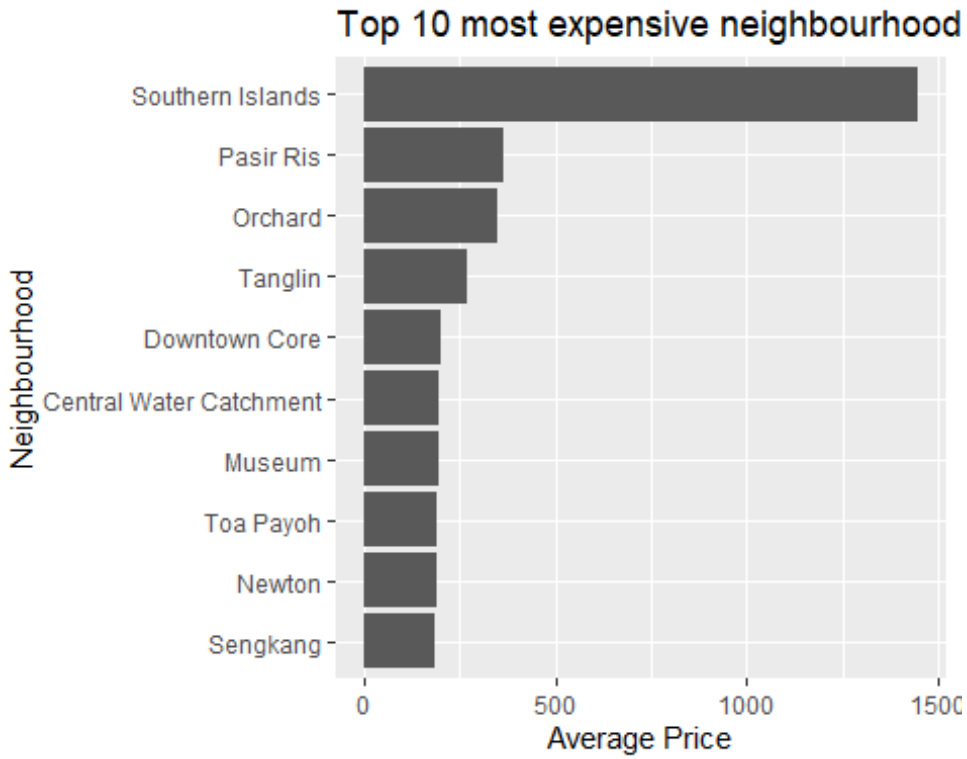
- **Bukit Timah**

Based on the results,we attempt to findout the reason of why Tuas having such high average price. After referring to the freq_area and airbnb dataframe, we found that the reason could be Tuas has only one listing in the neighbourhood and it is a private luxury pent house condo unit.The same goes to Marina South,Pioneer,Changi,Sungei Kadut,Mandai and Western Water Catchment which only have one to maximum two listings.Thus,this analysis could become bias and not accurate.Given example scenario, if a neighbourhood has less number of listings but with extremely high price or low price,its average value will be computed depending on certain one or two listings only.

In order to reduce biasness in our analysis, we set a threshold to the frequency of listings in the neighbourhood.By doing this, we ***filter out neighbourhood with listings frequency > 10***.

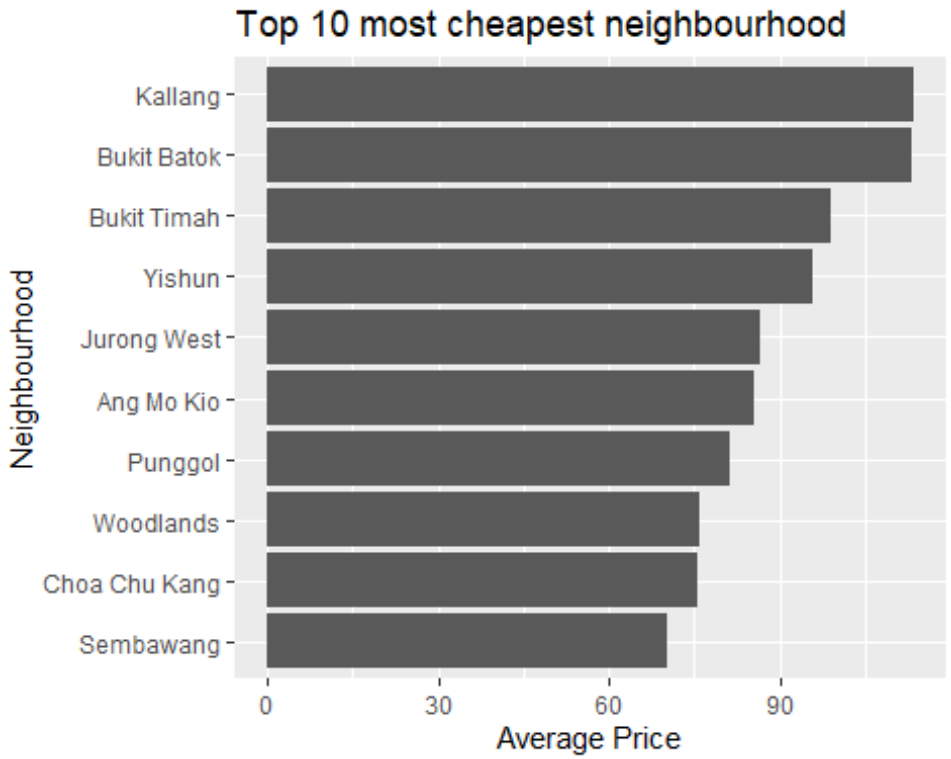
```
#compute average price per neighbourhood
avg_price_neighbourhood <- airbnb %>% group_by(neighbourhood)%>%
summarise(avg_price=mean(price))

## `summarise()` ungrouping output (override with `.groups` argument)

#construct frequency table of neighbourhood with average price
neighbourhood_df <- data.frame(Frequency =
table(airbnb$neighbourhood),avg_price_neighbourhood$avg_price)
colnames(neighbourhood_df)[1] <- "neighbourhood"
```

```r
colnames(neighbourhood_df)[2] <- "frequency"
colnames(neighbourhood_df)[3] <- "avg_price"

#Top 10 most expensive neighbourhood
#filter only listings of neighbouhood with n>10 to reduce bias
sorted_avg_price_exp <- neighbourhood_df %>% filter(frequency>10) %>%
arrange(desc(avg_price))

ggplot(sorted_avg_price_exp[1:10,],aes(x=reorder(neighbourhood,avg_price),
y=avg_price)) +
  geom_bar(stat='identity') + ggtitle('Top 10 most expensive neighbourhood')+
  xlab("Neighbourhood") +
  ylab("Average Price") +
  coord_flip()
```




```r
#Top 10 cheapest neighbourhood
#filter only listings of neighbouhood with n>10 to reduce bias
sorted_avg_price_cheap <- neighbourhood_df %>% filter(frequency>10) %>%
arrange(avg_price)


ggplot(sorted_avg_price_cheap[1:10,],aes(reorder(neighbourhood,avg_price),
y=avg_price)) +
  geom_bar(stat='identity') + ggtitle('Top 10 most cheapest neighbourhood')+
  xlab("Neighbourhood") +
```

```
  ylab("Average Price") +
  coord_flip()
```



The new results obtained are as follow:

The list of ***Top 10 Most Expensive Neighbourhoods*** :

Central Region：

- **Southern Islands**
- **Orchard**
- **Tanglin**
- **Downtown Core**
- **Musuem**
- **Toa Payoh**
- **Newton**

East Region：

- **Pasir Ris**

North Region:

- **Central Water Catchment**

North-East Region:

- **Sengkang**

Based on the analysis,renting rooms at the neighbourhood of Central Region such as Marina South,Southern Islands, Orchard, Downtown Core could be costly as these areas are more exclusive and affluent with shopping malls,restaurants and tourist attractions such as Merlion Park,Marina Bay Sands,Sentosa and many more.

The list of ***Top 10 Most Cheapest Neighbourhoods*** are as follow:

North Region:

- **Sembawang**
- **Woodlands**
- **Yishun**

West Region:

- **Choa Chu Kang**
- **Jurong West**
- **Bukit Batok**

North-East Region:

- **Punggol**
- **Ang Mo Kio**

Central Region:

- **Bukit Timah**
- **Kallang**

The price of listings within these neighbourhood are far more cheaper and affordable with average price ranges from around SGD70 to below SGD100. This could be due to most of them are residential town so its demand is not as high as in the tourist attractions.

Analysis Question 3:

Text Analysis on Singapore Airbnb reviews

```
library(rvest)

## Warning: package 'rvest' was built under R version 4.0.3

## Loading required package: xml2

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.0.3

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

```
## v tibble  3.0.3     v purrr   0.3.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------
--------------------- tidyverse_conflicts() --
## x glue::collapse()        masks dplyr::collapse()
## x dplyr::filter()         masks stats::filter()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
## x purrr::pluck()          masks rvest::pluck()

library(tidytext)
library(textstem)

## Loading required package: koRpus.lang.en

## Loading required package: koRpus

## Loading required package: sylly

## For information on available language packages for 'koRpus', run
##
##   available.koRpus.lang()
##
## and see ?install.koRpus.lang()

##
## Attaching package: 'koRpus'

## The following object is masked from 'package:readr':
##
##     tokenize

library(dplyr)
library(stringr)
library(tidyr)
library(ggplot2)
```

Before conducting text analysis,data cleaning steps were perfomed such as text tokenization,lowercase conversion,word lemmatization, numbers and stopwords filteration. Not to forget,non-alphanumeric characters such as foreign language that R cannot recognize and blank values are removed as well.Next,the cleaned dataset was used to perform further text analysis - Wordcloud and sentiment analysis.

```
#loading dataset
airbnb_reviews <- read.csv("C:/Users/user/Documents/R project/Singapore
Airbnb/reviews.csv")

#preview the dataset records
head(airbnb_reviews, 10)
```

```
##    listing_id       id       date reviewer_id reviewer_name
## 1       49091  8243238 2013-10-21     8557223         Jared
## 2       50646 11909864 2014-04-18     1356099         James
## 3       50646 13823948 2014-06-05    15222393         Welli
## 4       50646 15117222 2014-07-02     5543172         Cyril
## 5       50646 15426462 2014-07-08      817532          Jake
## 6       50646 15552912 2014-07-11    10942382         Subba
## 7       50646 15884470 2014-07-17    17569265        Claire
## 8       50646 16123989 2014-07-22    17188672          Hana
## 9       50646 16632638 2014-07-30    18067306           Liz
## 10      50646 16729657 2014-08-01     9211315       Derrick
##
comments
## 1
Fran was absolutely gracious and welcoming. Made my stay a great experience.
Would definitely recommend this cozy and peaceful place to anyone.
## 2
A comfortable room in a smart condo development. Everything was kept very
clean and I had the use of my own bathroom. Sujatha and her husband are great
hosts - very friendly and accommodating. I'll be staying here again.
## 3  Stayed over at Sujatha's house for 3 good nights with my boyfriend.
Sujatha and her husband are great hosts, very welcoming and friendly. The
room is comfortable and clean. I'm happy to have my own bathroom as i'm
particular with shared bathroom. \nThe location is accessible. A few minutes
walk from the house to nearest bus stop which can bring you to town.\nGood
place, good hosts, good price.\nHighly recommended!
## 4
It's been a lovely stay at Sujatha's. The room is clean and the location is
just perfect for a stop-over in Singapore. I really enjoyed relaxing at the
swimming pool after spending most of the day in the city. Thank you Sujatha.
## 5
We had a great experience. A nice place, an amazing complex and easy access
to public transit
## 6
Quiet condo. Comfortable stay and good location.
## 7
Nice room and friendly stay. Kindely and smiling family.
## 8
Suja and her husband are really nice, amazing, caring people.The room and
bathroom was clean. We truly enjoyed our stay.\nWill be back again next time!
## 9
Sujatha is a wonderful host and gives us a lot of help. The bedroom is cozy
with a good view of the mountain. Bus stop is just around the corner. Though
a little bit far from the entertainment hub of Singapore, Sujatha's condo is
worth all the compliments in the world!
## 10
A wonderful experience & highly recommended! Sujatha has a very comfortable
home and really made me feel welcome!! I am very grateful to Sujatha & her
husband, Devi, and I look forward to seeing you both again soon.  =))
```

```
#Data cleaning
#tokenize the text
#convert all the words to lowercase
#lemmatize the words
clean_review= airbnb_reviews%>%
  unnest_tokens(word,comments, token="words",to_lower=TRUE)%>%
  mutate(word=lemmatize_words(word,dictionary=lexicon::hash_lemmas))

#remove numbers
#no_numbers= filter(clean_review,is.na(as.numeric(gsub(",","",word))))
no_numbers = filter(clean_review,is.na(as.numeric(clean_review$word)))

## Warning in mask$eval_all_filter(dots, env_filter): NAs introduced by
coercion

#filter stopwords
no_stop_words= anti_join(no_numbers,stop_words,by ="word")

#remove all non-alphanumeric chracters
no_special_df <- as.data.frame(gsub("[^0-9A-Za-z///' ]","" ,
no_stop_words$word ,ignore.case = TRUE,))
colnames(no_special_df)[1] <- "word"

#remove blank values
no_blank_df <- as.data.frame(no_special_df[!apply(no_special_df == "", 1,
all), ])
colnames(no_blank_df)[1] <- "word"
```

WordCloud

Word cloud helps to visualize the most important and the most frequent words that are being mentioned in the reviews.

```
library("wordcloud")

## Loading required package: RColorBrewer

library("RColorBrewer")

#count the frequency of each word
word_freq=count(no_blank_df,word,sort = TRUE)

#wordcloud
wordcloud(word_freq$word,word_freq$n, min.freq=3, max.words=100,
          random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8,"Dark2"))
```

From the analysis, it is observed that words such as **stay**,**clean**,**location**,**host**,**recommend**,**comfortable**,**convenient**,**mrt** and etc appeared most oftenly.

From here,we can tell that most of the listings in the Singapore Airbnb are adhered to these standard practices.Not surprising that all these words are being stand out in the word cloud as these are the the most important criteria and elements when people are looking for a stay at Airbnb.

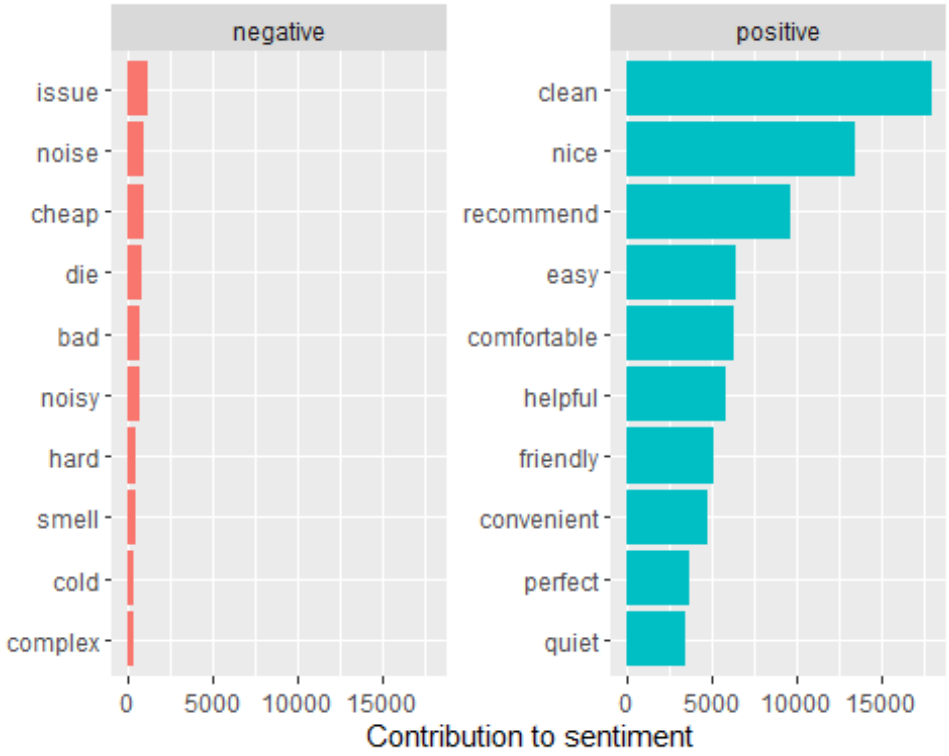
Sentiment Analysis

Moving on to the sentiment analysis,it help us to determine the emotional tone behind a series of words as well as gain an understanding of the the attitudes,opinions and emotions expressed within the airbnb reviews.

```
#get specific sentiment lexicons
bing= get_sentiments("bing")

#return rows in no_blank_df that have a match in bing
bing_words= inner_join(no_blank_df,bing,by="word")

#count the number of words in each sentiment group
count_sentiment= bing_words %>%
  group_by(word,sentiment) %>%
  count(word,sort=TRUE)
```

```r
#word count frequnecy plot
count_sentiment %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(y = "Contribution to sentiment",
       x = NULL) +
  coord_flip()
```

```
## Selecting by n
```



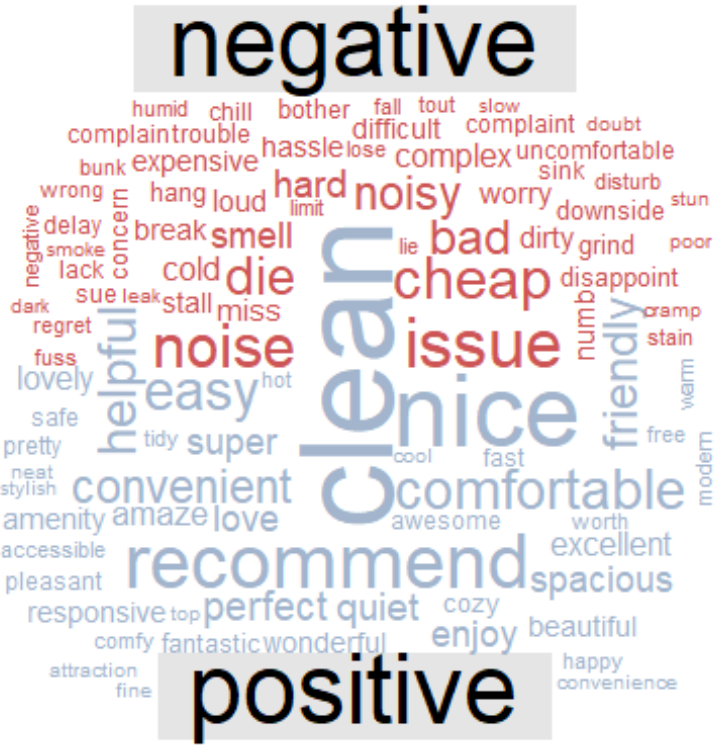

```r
#inspect certain words to understand its details
inspect=filter(airbnb_reviews,str_detect(comments,"issue"))

#comparisoin word cloud
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
no_blank_df %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("indianred3","lightsteelblue3"),
                   max.words = 100)
```

```
## Joining, by = "word"
```



The **positive words** are ***clean***,***nice***,***recommend***,***easy***,***friendly***,***convenient*** which express the feeling of the guests that they were ***satisfied*** with the facilities,environment and friendly attitude of the hosts which prompt them to leave a good review and recommend the listings to others.

On the other hand,**negative words** such as ***noisy***,***smell***,***bad***,***cheap***,***uncomfortable*** were mentioned in the reviews.The guests feel ***disappointed*** due to some factors,which we made assumption that they were disturbed by the noise as the rooms are not designed to be soundproof.Hosts may take attention to this matter and further improve it to attract and retain more guests.We also assumed that most of the guests will have a thought of they are facing all these bad experience because of the cheap price so that **cheap** word has a negative sentiment in the review.

Bigrams

```
library(igraph)
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:purrr':
##
##     compose, simplify

## The following object is masked from 'package:tidyr':
##
##     crossing

## The following object is masked from 'package:tibble':
##
##     as_data_frame

## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union

library(ggraph)

## Warning: package 'ggraph' was built under R version 4.0.3

#Data cleaning
#tokenize the text
#convert all the words to lowercase
#lemmatize the words
clean_review_bigram= airbnb_reviews %>%
  unnest_tokens(word,comments,token="ngrams",n=2,to_lower=TRUE)%>%
  mutate(word=lemmatize_words(word,dictionary = lexicon::hash_lemmas))

#remove numbers
bigrams_no_numbers =
filter(clean_review_bigram,is.na(as.numeric(clean_review_bigram$word)))

## Warning in mask$eval_all_filter(dots, env_filter): NAs introduced by
coercion

#remove all non-alphanumeric chracters
bigrams_no_special <- as.data.frame(gsub("[^0-9A-Za-z///' ]","" ,
bigrams_no_numbers$word ,ignore.case = TRUE,))
colnames(bigrams_no_special)[1] <- "word"

#remove blank values
bigrams_no_blank <-
as.data.frame(bigrams_no_special[!apply(bigrams_no_special == "", 1, all), ])
colnames(bigrams_no_blank)[1] <- "word"
```

```r
#seperate into word1 and word2
bigrams_separated = bigrams_no_blank %>%
  separate(word,c("word1","word2"),sep = " ")

my_stop_words= tibble (word=c("singapore","airbnb"),
                               lexicon="custom")

#connect both stop words data frames
all_stop_words= bind_rows(stop_words,my_stop_words)

bigrams_filtered = bigrams_separated %>%
  filter(!word1 %in% all_stop_words$word) %>%
  filter(!word2 %in% all_stop_words$word) %>%
  filter(is.na(as.numeric(gsub(",","",word1)))) %>%
  filter(is.na(as.numeric(gsub(",","",word2))))
```

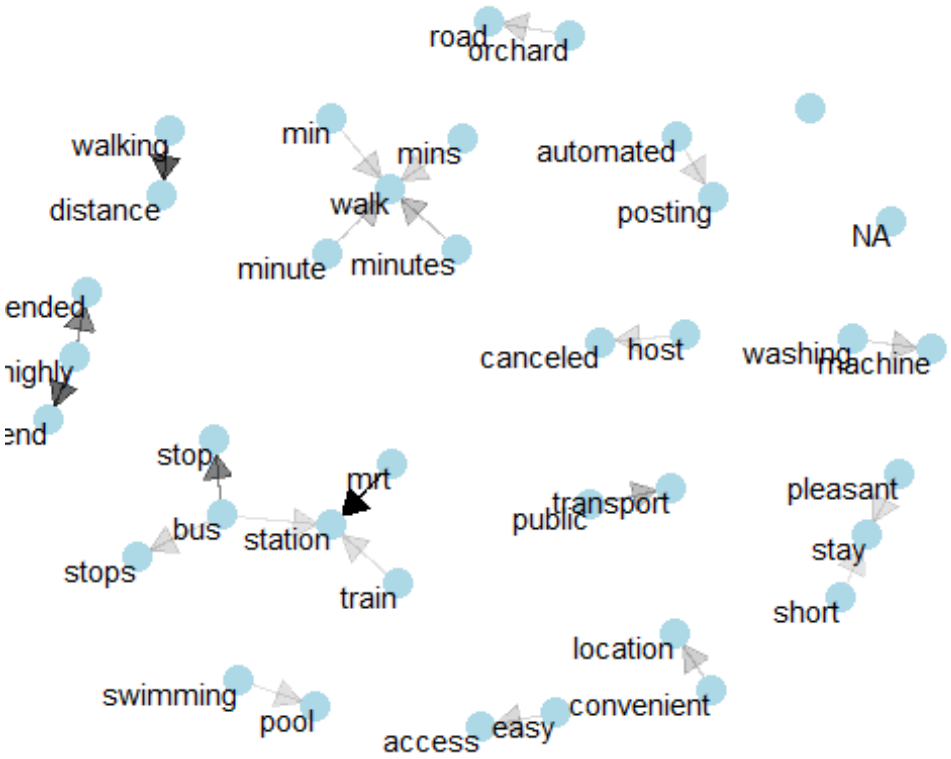
```
## Warning in mask$eval_all_filter(dots, env_filter): NAs introduced by
coercion
```

```
## Warning in mask$eval_all_filter(dots, env_filter): NAs introduced by
coercion
```

```r
# count of bigram
bigram_counts <- bigrams_filtered %>%
  count(word1, word2, sort = TRUE)

# filter for only relatively common combinations
bigram_graph <- bigram_counts %>%
  filter(n > 500) %>%
  graph_from_data_frame()
```

```
## Warning in graph_from_data_frame(.): In `d' `NA' elements were replaced
with
## string "NA"
```

```r
a <- grid::arrow(type = "closed", length = unit(.15, "inches"))

ggraph(bigram_graph, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1) +
  theme_void()
```



The *unnest_tokens* function is used to tokenize the text into consecutive sequences of words,called n-grams which enabled us to examine what words tend to appear after others. We are interested in visualizing a network of bigram and all of the relationships among words simultaneously.Based on the common bigrams,it showS the bigrams that occurred more than 500 times and where neither word was a stop word.We observed pairs that form common short phrases such as **mrt station**,**bus stop**, **pleasant stay**,**public transport**.

D.Conclusion and Lesson Learned

In a nutshell,the descriptive analysis outlines the average market price of Singapore Airbnb listings which can act as useful guildlines for the hosts to determine the rental rates.Furthermore,the text analysis highlights few key requirements which guests looking for when they are booking a room.The hosts can now better understanding the guests' needs and make necessarily improvement based on the bad reviews in order to stay ahead their competitors.Besides,Singapore has many different Airbnb listings options that suits every budget level.Thus,do not be deterred by its reputation as being an expensive city.This analysis can be further improved by examining the factors that affect the price and performing predictive analysis.

The lesson learned from this assignment is we should pay more attention to details of variables during data exploration phases.Sometimes,we tend to neglect some details and caused our analysis to be biased and inaccurate.By then,it reminded us to look and think of different perspective when dealing with the variables.Besides,the codes could be improved

and modified into a simpler way by making use of functions and packages to increase its readability and efficiency .

E.Individual Reflection- Chan Wei Wei 16052748

This assignment has definitely help to concrete my understanding on the web analytics concepts learned.Starting from topic research,dataset exploration,data analysis to findings interpretation, I learned from the trial and error process upon completing this project.Now,I become more familiar with the syntax and structure of R programming language.In my opinion,R tools has definitely make the life of data analyst a lot more easier with the vast numbers of built-in packages and robust graphical capabilities.As a result,more time can be spend and focus on producing high-quality analysis tasks.For example,I had made use of the dplyr and ggplot packages for data manipulation and plotting purposes during the analysis.

The biggest challenge I faced was when performing data cleaning part for Airbnb reviews.I was stucked at removing the special characters and NULL value in the reviews as I am not familiar with the functions in R.Fortunately, I am able to solve the problems with the help from the stackoverflow.This project also inspired me to continue working on other area of analysis to gain further insights.

Individual Reflection- Chan Wei Chee 16052755

Upon completing this assignment,I am able to apply the web analysis techniques learned from the lab exercises and better relate the knowledge to solve real life industry problems. Also,conducting analysis tasks using R is not as difficult as i think since there are vast numbers of built-in libraries and packages available to ease the analysis tasks.I was impressed by the syntax simplicity of the R functions and packages.For instance ,by installing tidyverse packages,I am able to form a dataframe and generate plots and graphs with just few lines of codes.

The exciting part was the time when we were able to derive some insights from the unstructured datasets.The analysis result also improved my understanding on the demand level and property listings at Singapore which might be heplful in the future.I will continue honing R programming as a good analyst needs to be proficient in R in order to excel in the domain of data science.