Homework 1

105304028_統計四_方品謙

線上連續劇觀看資料

安裝package

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.0     ✓ purrr   0.3.3
## ✓ tibble  2.1.3     ✓ dplyr   0.8.4
## ✓ tidyr   1.0.2     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0
```

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

```
library(readr)
```

匯入資料

首先必須設定工作目錄(setwd("工作目錄"))，再讀入資料

```
watch.table <- read_csv("watch_table.csv")
```

```
## Parsed with column specification:
## cols(
##   watch_id = col_character(),
##   time_stamp = col_character(),
##   user_id = col_character(),
##   drama_id = col_character(),
##   device = col_character()
## )
```

```
user.table <- read_csv("user_table.csv")
```

```
## Parsed with column specification:
## cols(
##   user_id = col_character(),
##   user_name = col_character(),
##   gender = col_character(),
##   age = col_double(),
##   location = col_character(),
##   payment = col_double()
## )
```

```
drama.table <- read_csv("drama_table.csv")
```

```
## Parsed with column specification:
## cols(
##   drama_id = col_character(),
##   drama_name = col_character(),
##   area = col_character(),
##   actors = col_character(),
##   description = col_character()
## )
```

1.將 watch.table 與其他兩個報表合併為full.table

```
full.table <- watch.table %>%
  left_join(user.table, by = "user_id") %>%
  left_join(drama.table , by = "drama_id")
```

2.計算每部劇男生、女生觀看次數

```
full.table %>% group_by(gender) %>%
  select( gender) %>%
  summarize(nmuber = n() )
```

gender <chr>	nmuber <int>
female	18
male	21

2 rows

3.針對用Android系統的客戶進行分析

```
full.table %>%
  group_by(user_name , age ,location) %>%
  filter(device == "Android") %>%
  summarize(male = table(unique(gender))[1],
            female = table(unique(gender))[2])
```

user_name <chr>	age <dbl>	location <chr>	male <int>	female <int>
Alex Chu	25	Taipei	1	*NA*
Sandy Wu	37	Hsinchu	1	*NA*

2 rows

使用Android系統的客戶只有兩名，全為男性，分別為25歲及37歲。

4.針對台北男性客戶進行分析

```
full.table %>%
  filter(location == "Taipei" & gender == "male" ) %>%
  group_by(user_name, device , age ) %>%  summarize(gender = unique(gender))
```

user_name <chr>	device <chr>	age <dbl>	gender <chr>
Alex Chu	Android	25	male
Alex Chu	Chrome	25	male
Andy Liu	iOS	26	male
Eric Chou	Chrome	28	male
Jacky Wu	iOS	24	male
Jay Chou	Chrome	29	male
Tom Gu	IE	29	male

7 rows

由上表可以看到，台北男性最常使用Chrome為device，iOS為其次。而年齡介於24-29歲之間，對比所有觀看客戶來說相對年輕。

kaggle上2019紐約Airbnb的資料

匯入資料

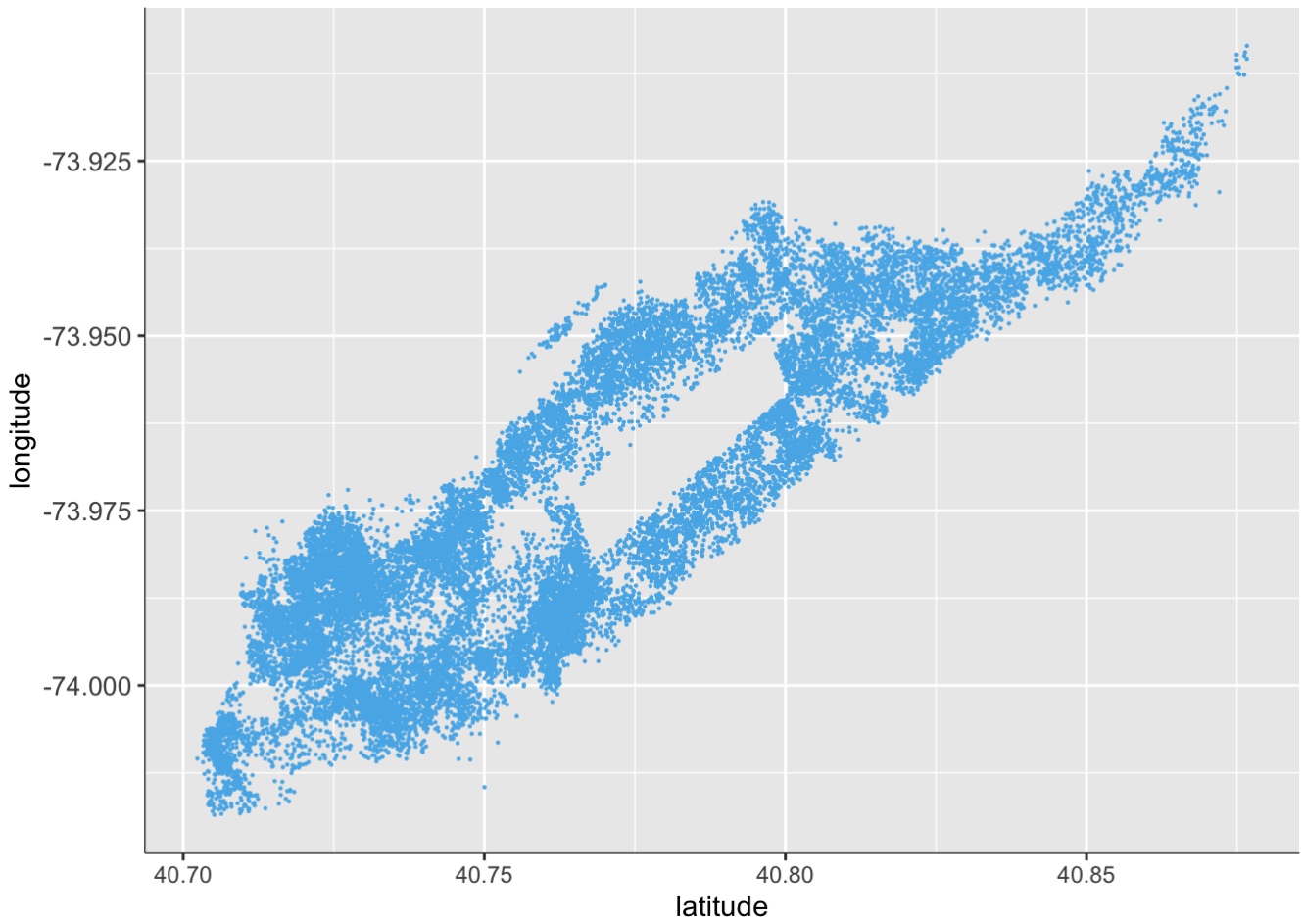
首先必須設定工作目錄(setwd("工作目錄"))，再讀入資料

```
air_bnb <- read_csv("AB_NYC_2019.csv")
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   name = col_character(),
##   host_id = col_double(),
##   host_name = col_character(),
##   neighbourhood_group = col_character(),
##   neighbourhood = col_character(),
##   latitude = col_double(),
##   longitude = col_double(),
##   room_type = col_character(),
##   price = col_double(),
##   minimum_nights = col_double(),
##   number_of_reviews = col_double(),
##   last_review = col_date(format = ""),
##   reviews_per_month = col_double(),
##   calculated_host_listings_count = col_double(),
##   availability_365 = col_double()
## )
```

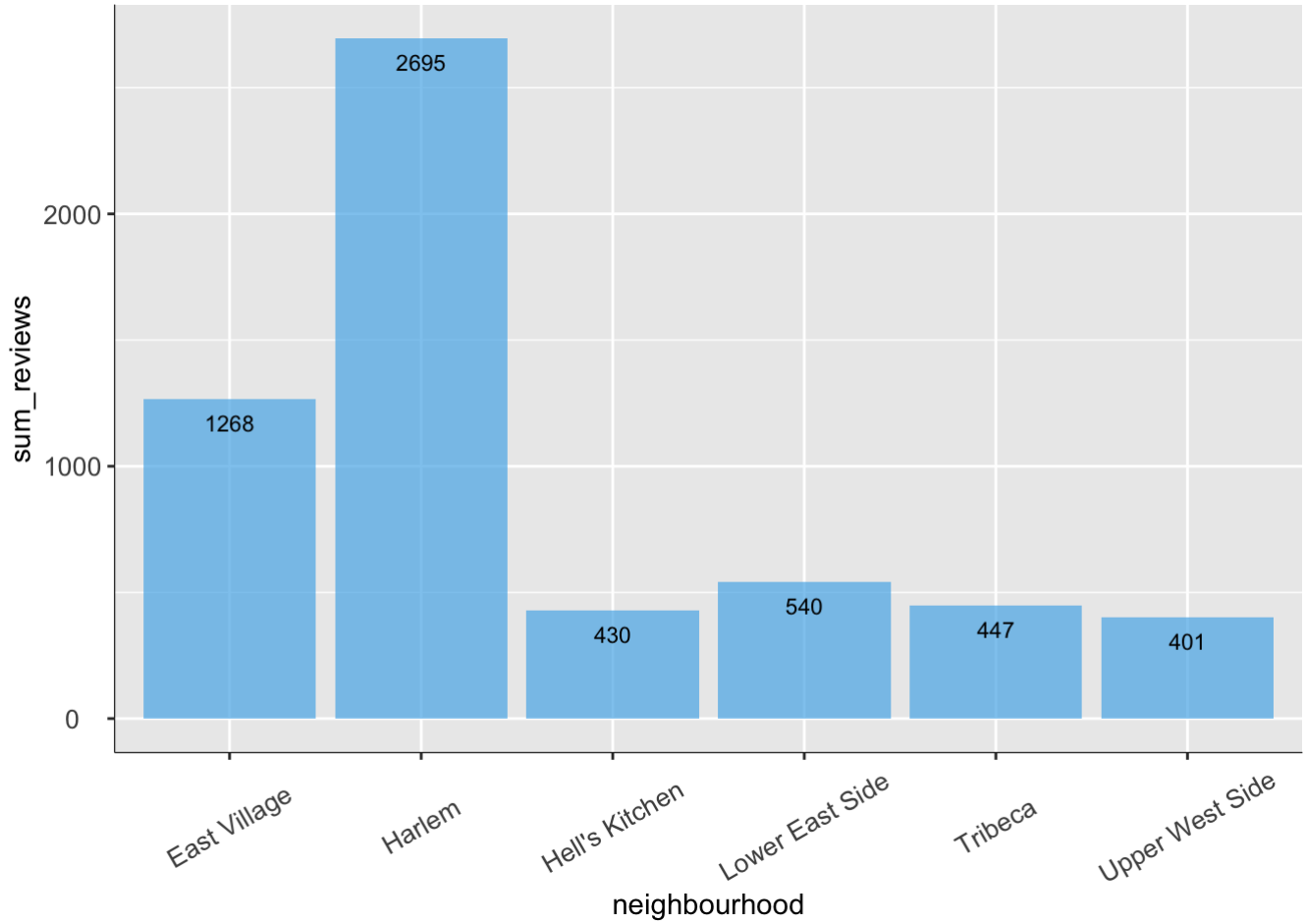
1.利用ggplot畫經緯度的scatter plot

```
air_bnb %>% filter( neighbourhood_group == "Manhattan" ) %>%
  ggplot( air_bnb , mapping = aes(x = latitude , y = longitude)) + geom_point(size =
0.1, color = "#56B4E9")+
  theme(axis.text.y = element_text(size = 10 , vjust = 0.5, hjust = 0.5))+
  theme(axis.line = element_line(size=0.2, colour = "black"))
```



2.針對曼哈頓資畫bar chart

```
air_bnb %>%
  filter(neighbourhood_group == "Manhattan"  & number_of_reviews >=400) %>%
  group_by(neighbourhood) %>%
  summarise(sum_reviews = sum(number_of_reviews)) %>%
ggplot(air_bnb, mapping = aes(x = neighbourhood , y = sum_reviews)) + geom_bar(stat=
"identity" , fill = "#56B4E9" , alpha = 0.7) +
  geom_text(stat="identity",aes(label=sum_reviews),vjust=2, color=I("#000000"),size=3
)+
  theme(axis.text.x = element_text(size = 10 , vjust = 0.5, hjust = 0.5 , angle = 30
))+
  theme(axis.text.y = element_text(size = 10 , vjust = 0.5, hjust = 0.5))+
  theme(axis.line = element_line(size=0.2, colour = "black"))
```



3.擁有最多number_of_reviews的neighbourhood

```
air_bnb %>%
  filter(neighbourhood_group == "Manhattan"  & number_of_reviews >=400) %>%
  group_by(neighbourhood) %>%
  summarise(sum_reviews = sum(number_of_reviews)) %>%
  select(neighbourhood,sum_reviews) %>%
  filter(rank(desc(sum_reviews)) == 1)
```

neighbourhood	sum_reviews
<chr>	<dbl>
Harlem	2695

1 row

4.EDA分析

去除掉NA值

```
Harlem <- air_bnb %>%
              filter(number_of_reviews >=400 & neighbourhood == "Harlem")
table(is.na(Harlem))
```

```
##
## FALSE
##    80
```

```
Harlem <- na.omit(Harlem)
table(is.na(Harlem))
```

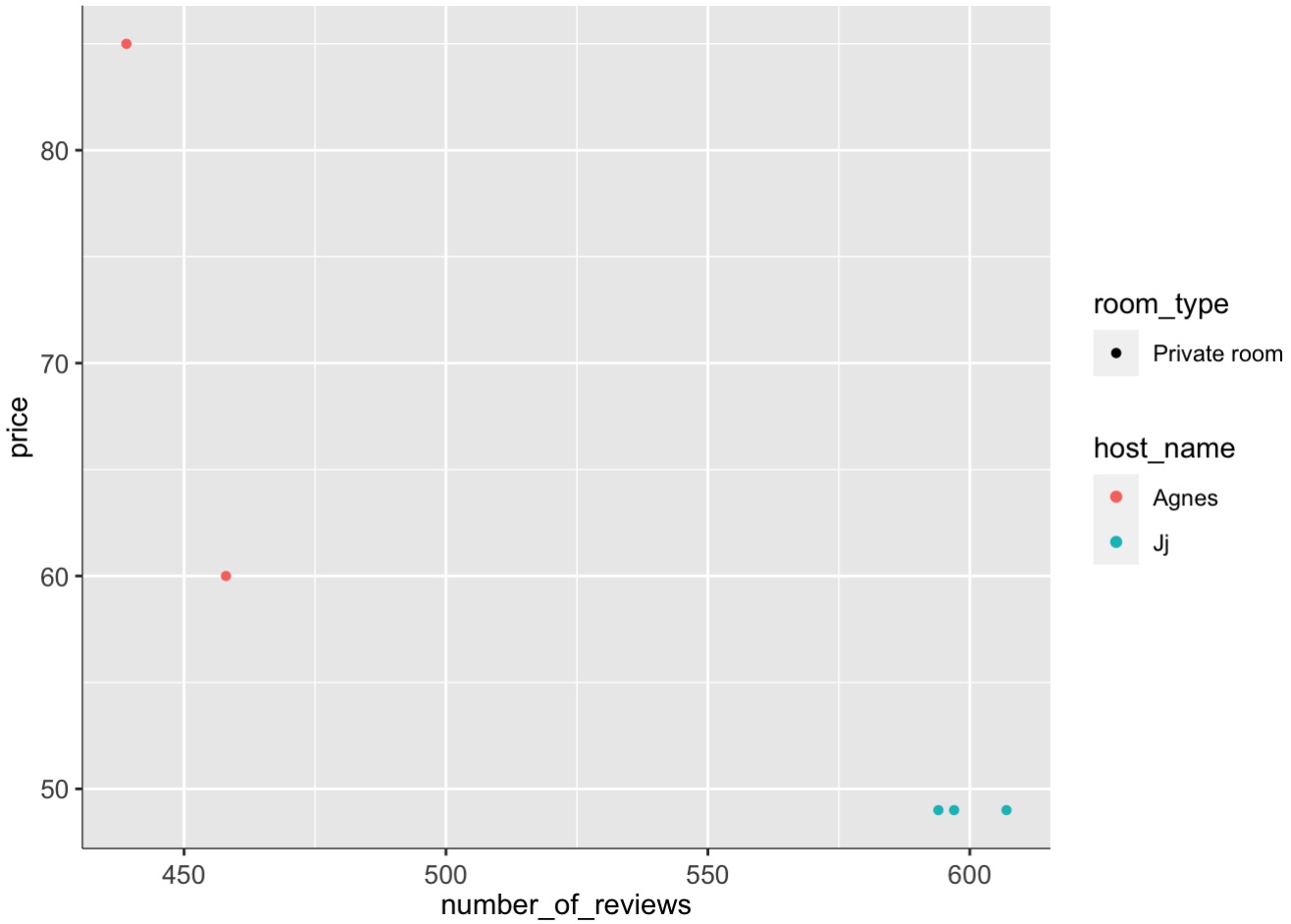
```
##
## FALSE
##    80
```

EDA分析

```
str(Harlem)
```

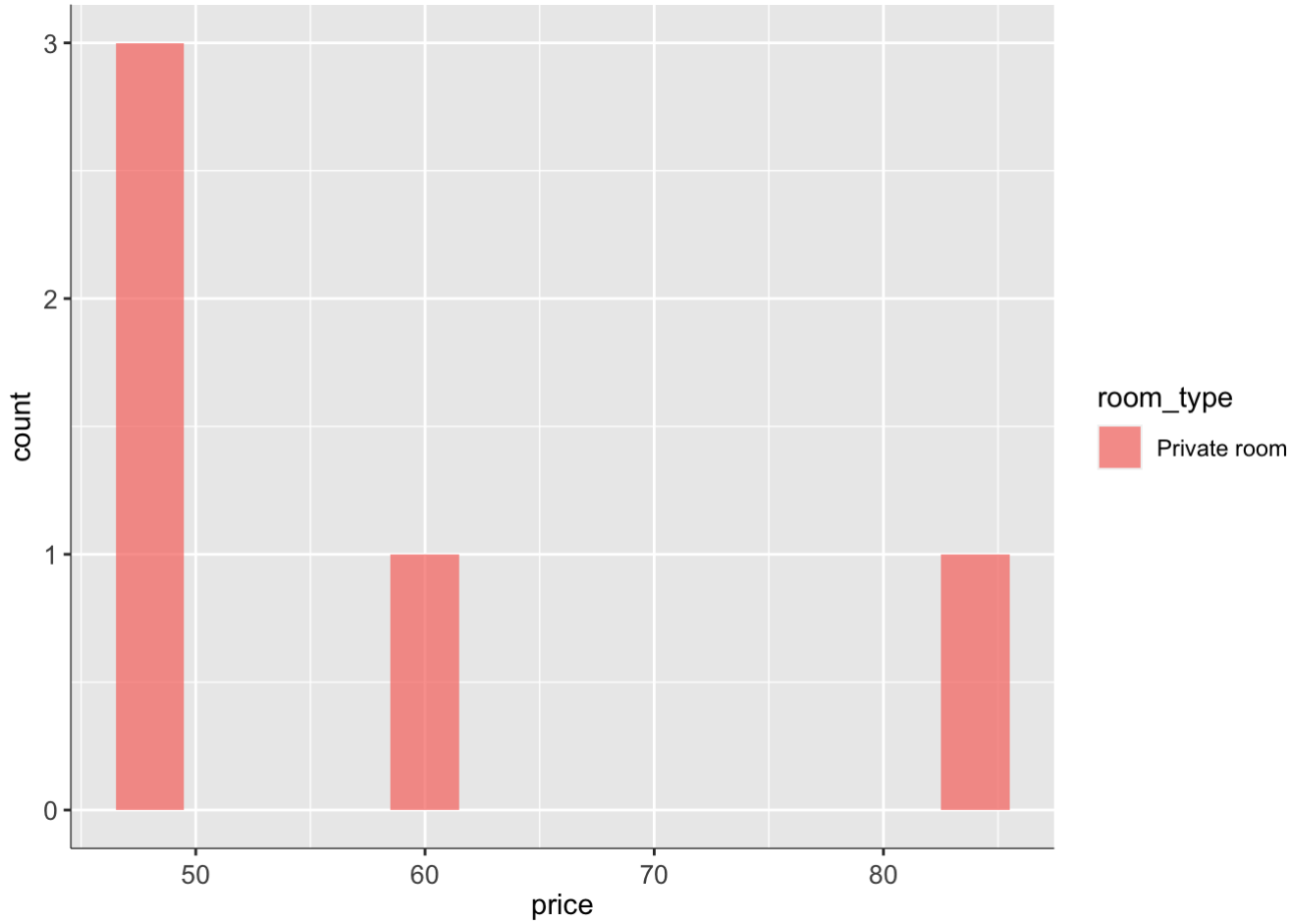
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    5 obs. of  16 variables:
##  $ id                             : num  58059 476983 891117 903947 903972
##  $ name                           : chr  "PRIVATE Room on Historic Sugar Hill" "PRI
VATE Room in Spacious, Quiet Apt" "Private Bedroom in Manhattan" "Beautiful Bedroom i
n Manhattan" ...
##  $ host_id                        : num  277379 277379 4734398 4734398 4734398
##  $ host_name                      : chr  "Agnes" "Agnes" "Jj" "Jj" ...
##  $ neighbourhood_group            : chr  "Manhattan" "Manhattan" "Manhattan" "Manha
ttan" ...
##  $ neighbourhood                  : chr  "Harlem" "Harlem" "Harlem" "Harlem" ...
##  $ latitude                       : num  40.8 40.8 40.8 40.8 40.8
##  $ longitude                      : num  -73.9 -73.9 -73.9 -73.9 -73.9
##  $ room_type                      : chr  "Private room" "Private room" "Private roo
m" "Private room" ...
##  $ price                          : num  60 85 49 49 49
##  $ minimum_nights                 : num  1 1 1 1 1
##  $ number_of_reviews              : num  458 439 594 597 607
##  $ last_review                    : Date, format: "2019-07-03" "2019-07-05" ...
##  $ reviews_per_month              : num  4.58 5.12 7.57 7.72 7.75
##  $ calculated_host_listings_count: num  2 2 3 3 3
##  $ availability_365               : num  258 238 339 342 293
```

```
ggplot(Harlem , aes(x = number_of_reviews, y = price, color = host_name, shape = room
_type)) +
  geom_point(size=1.5)+
  theme(axis.text.x = element_text(size = 10 , vjust = 0.5, hjust = 0.5))+
  theme(axis.text.y = element_text(size = 10 , vjust = 0.5, hjust = 0.5))+
  theme(axis.line = element_line(size=0.2, colour = "black"))
```



number_of_reviews大於400且neighbourhood為Harlem的共有五間房子，評論最多的607則。但價格最高的反而是評論數最少的房子。 而評論最多的三間皆是Ji的房子，其他則是Agnes的。

```
ggplot(Harlem , aes(x = price, fill = room_type)) + geom_histogram(alpha = 0.7 , binw
idth=3)+
  theme(axis.text.x = element_text(size = 10 , vjust = 0.5, hjust = 0.5))+
  theme(axis.text.y = element_text(size = 10 , vjust = 0.5, hjust = 0.5))+
  theme(axis.line = element_line(size=0.2, colour = "black"))
```



最高房價為85元，類型為Private room。最低房價為49元，類型為Private room。