商業分析 HW1

105305072 企管四 許惠甄

一、

1. 將 watch.table 與其他兩個報表合併為 full.table。

```
## 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")

full.table
```

2. 分析 full.table，計算每部劇男生、女生觀看次數 (group_by(), summarize(), length())

```
## 2. 分析full.table，計算每部劇男生、女生觀看次數
full.table %>%
  group_by(drama_name) %>%
  summarise(female_number = length(which(gender == "female")),
            male_number = length(which(gender == "male")))
```

呈現結果如下：

drama_name	female_number	male_number
<chr>	<int>	<int>
Bromance	1	2
Descendants Of The Sun	3	5
From 5 to 9	0	6
House of Cards	3	4
I Should Not Love You	3	0
She was Pretty	4	1
The Flame's Daughter	4	3

從表格中可看出《From 5 to 9》的男性觀看次數最多，而《She was Pretty》與《The Flame’ s Daughter》是女性觀看次數較多的兩部劇。

3. 找出用 Android 系統的，針對這類客戶進行分析（例如平均年紀、性別、來自哪裡）。

```
## 3. 找出用Android系統的，針對這類客戶進行分析。
full.table %>%
  filter(device == "Android") %>%
  summarise(Avg_age = mean(age),
            total_number = n())
```

呈現結果如下：

	Avg_age	total_number
	<dbl>	<int>
1	28.4	7

從表格中可看出使用 Android 系統的平均年齡為 28.4 歲，佔原先 39 筆資料中的 7 筆。

```r
full.table %>%
  group_by(gender) %>%
  filter(device == "Android") %>%
  summarise(gender_distribute = n())

full.table %>%
  group_by(drama_name) %>%
  filter(device == "Android") %>%
  summarise(drama_distribute = n())

full.table %>%
  group_by(location) %>%
  filter(device == "Android") %>%
  summarise(location_distribute = n())

full.table %>%
  group_by(user_name) %>%
  filter(device == "Android") %>%
  summarise(user_distribute = n())
```

```
# A tibble: 1 x 2
  gender gender_distribute
  <chr>              <int>
1 male                   7
# A tibble: 2 x 2
  drama_name     drama_distribute
  <chr>                     <int>
1 From 5 to 9                   5
2 House of Cards                2
# A tibble: 2 x 2
  location location_distribute
  <chr>                  <int>
1 Hsinchu                    2
2 Taipei                     5
# A tibble: 2 x 2
  user_name user_distribute
  <chr>               <int>
1 Alex Chu                5
2 Sandy Wu                2
```

從以上表格中可看出使用 Android 系統的共有兩位男性，一共觀看《From 5 to 9》5 次及《House of Cards》2 次，居住地區為新竹及台北。

4. 針對台北男性這類客戶進行分析。

```r
## 4. 針對台北男性這類客戶進行分析。
full.table %>%
  filter(location == "Taipei" & gender == "male") %>%
  summarise(Avg_age = mean(age),
            total_number = n())
```

呈現結果如下：

```
# A tibble: 1 x 2
  Avg_age total_number
    <dbl>        <int>
1    26.1           17
```

從表格中可看出台北男性觀看者的平均年齡為 26.1 歲，佔原先 39 筆資料中的 17 筆。

```r
full.table %>%
  group_by(drama_name) %>%
  filter(location == "Taipei" & gender == "male") %>%
  summarise(drama_distribute = n())
```

呈現結果如下：

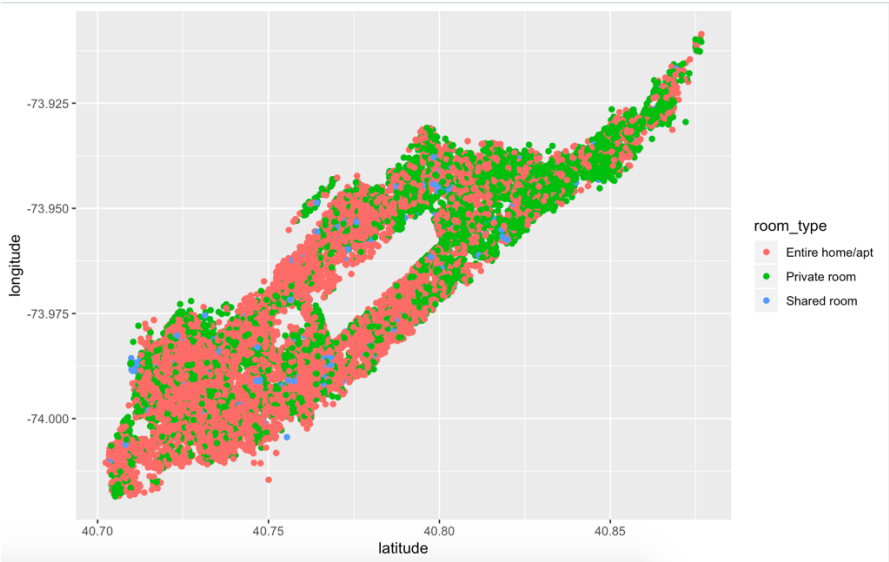
```
# A tibble: 5 x 2
  drama_name             drama_distribute
  <chr>                             <int>
1 Bromance                              2
2 Descendants Of The Sun                5
3 From 5 to 9                           6
4 She was Pretty                        1
5 The Flame's Daughter                  3
```

從表格中可看出台北男性觀看次數最多的影集是《From 5 to 9》，次數最少的是《She was Pretty》。

二、

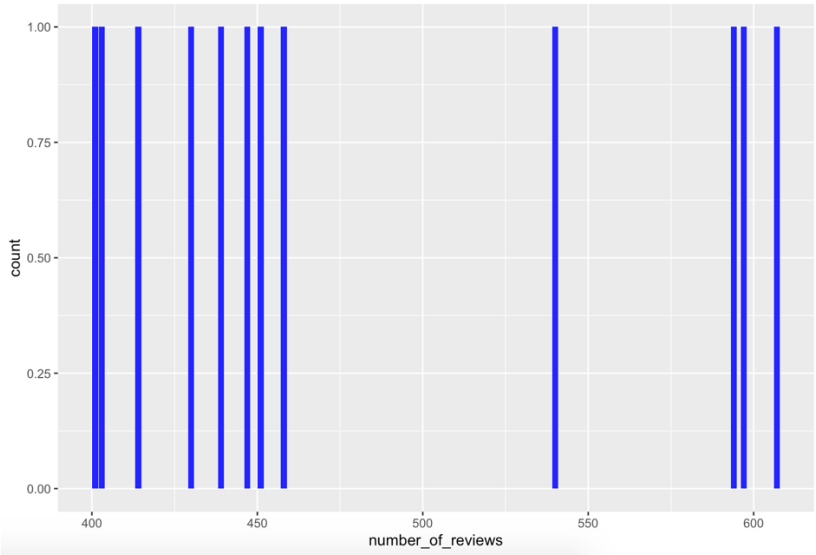
1. 找出 neighbourhood_group == "Manhattan"的資料，利用 ggplot 畫經緯度的 scatter plot。

```
## 1. 找出 neighbourhood_group == "Manhattan"的資料，利用ggplot畫經緯度的scatter plot。
abnyc.table %>%
  filter(neighbourhood_group == "Manhattan") %>%
  ggplot(aes(x=latitude, y=longitude, color=room_type)) +
  geom_point()
```



2. 針對曼哈頓資料，對 number_of_reviews >=400 的畫 bar chart。

```
## 2. 針對曼哈頓資料，對number_of_reviews >=400的畫bar chart。
abnyc.table %>%
  filter(neighbourhood_group == "Manhattan" & number_of_reviews >= 400) %>%
  ggplot(aes(x=number_of_reviews)) +
  geom_bar(fill="blue")
```



3. 針對曼哈頓資料，number_of_reviews >=400 的中，哪個 neighbourhood 擁有最多 number_of_reviews。

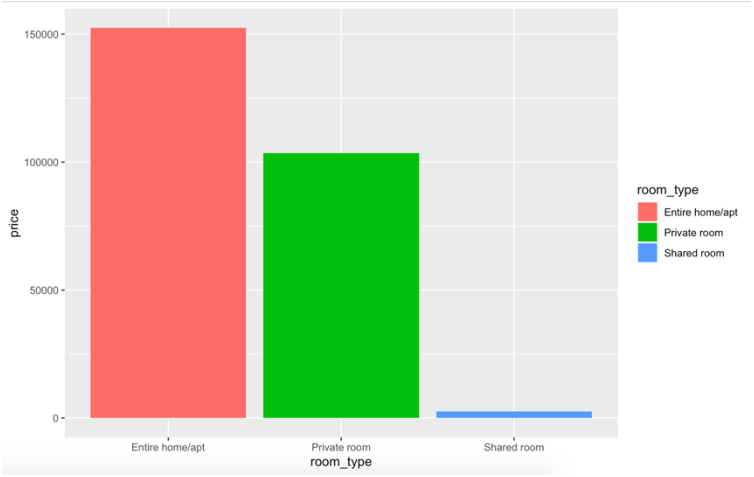
```
## 3. 針對曼哈頓資料，number_of_reviews >=400的中，哪個neighbourhood擁有最多number_of_reviews。
k <- abnyc.table %>%
  filter(neighbourhood_group == "Manhattan" & number_of_reviews >= 400)
i_max <- which.max(k$number_of_reviews)
k$neighbourhood[i_max]
```

結果顯示「Harlem」區域的房子有最高的 Review 次數。

4. 建立一筆新資料，將 3.找出的 neighbourhood 篩選出來，去除掉 NA 值後，進行 EDA 分析，並簡單介紹最高房價及最低房價分別的類型。

```
## 4.建立一筆新資料，將3.找出的neighbourhood篩選出來，去除掉NA值後，進行EDA分析，並簡單介紹最高房價及最低房價分別的類型。
new.table <- na.omit(abnyc.table %>%
  filter(neighbourhood == "Harlem"))

new.table %>%
  group_by(room_type) %>%
  arrange(desc(price)) %>%
  ggplot(aes(x=room_type, y=price, fill=room_type)) +
  geom_bar(stat = "identity")
```



結果顯示最高房價是「Entire home/apt」類型，而最低房價是「Share room」類型。

```
library(tidyverse)
library(readr)

#一、

watch.table <- read_csv("watch_table.csv")
user.table <- read_csv("user_table.csv")
drama.table <- read_csv("drama_table.csv")

## 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")

full.table

## 2. 分析full.table，計算每部劇男生、女生觀看次數
full.table %>%
  group_by(drama_name) %>%
  summarise(female_number = length(which(gender == "female")),
            male_number = length(which(gender == "male")))

## 3. 找出用Android系統的，針對這類客戶進行分析。
full.table %>%
  filter(device == "Android") %>%
  summarise(Avg_age = mean(age),
            total_number = n())

full.table %>%
  group_by(gender) %>%
  filter(device == "Android") %>%
  summarise(gender_distribute = n())

full.table %>%
  group_by(drama_name) %>%
  filter(device == "Android") %>%
  summarise(drama_distribute = n())

full.table %>%
  group_by(location) %>%
  filter(device == "Android") %>%
  summarise(location_distribute = n())

full.table %>%
  group_by(user_name) %>%
  filter(device == "Android") %>%
  summarise(user_distribute = n())

## 4. 針對台北男性這類客戶進行分析。
full.table %>%
  filter(location == "Taipei" & gender == "male") %>%
  summarise(Avg_age = mean(age),
            total_number = n())
```

```
full.table %>%
  group_by(drama_name) %>%
  filter(location == "Taipei" & gender == "male") %>%
  summarise(drama_distribute = n())


#二、

abnyc.table <- read.csv("AB_NYC_2019.csv")

## 1. 找出 neighbourhood_group == "Manhattan"的資料，利用ggplot畫經緯度的
scatter plot。
abnyc.table %>%
  filter(neighbourhood_group == "Manhattan") %>%
  ggplot(aes(x=latitude, y=longitude, color=room_type)) +
  geom_point()

## 2. 針對曼哈頓資料，對number_of_reviews >=400的畫bar chart。
abnyc.table %>%
  filter(neighbourhood_group == "Manhattan" & number_of_reviews >= 400) %>%
  ggplot(aes(x=number_of_reviews)) +
  geom_bar(fill="blue")

## 3. 針對曼哈頓資料，number_of_reviews >=400的中，哪個neighbourhood擁有最多
number_of_reviews。
k <- abnyc.table %>%
  filter(neighbourhood_group == "Manhattan" & number_of_reviews >= 400)
i_max <- which.max(k$number_of_reviews)
k$neighbourhood[i_max]

## 4.建立一筆新資料，將3.找出的neighbourhood篩選出來，去除掉NA值後，進行EDA分析，並簡
單介紹最高房價及最低房價分別的類型。
new.table <- na.omit(abnyc.table %>%
  filter(neighbourhood == "Harlem"))

new.table %>%
  group_by(room_type) %>%
  arrange(desc(price)) %>%
  ggplot(aes(x=room_type, y=price, fill=room_type)) +
  geom_bar(stat = "identity")
```