智慧型共享旅宿定價系統

指導老師: 蔡智勇老師 . 黃登揚老師
第三組

專題報告大綱

Project Overview

1. Airbnb 背景與專題主題介紹
2. 資料來源及描述
3. 資料清洗
4. 機器學習模型分析
5. 應用與結論

the A team

INTRODUCTION

背景與專題主題介紹

Airbnb 背景與專題主題介紹

Sharing Economy

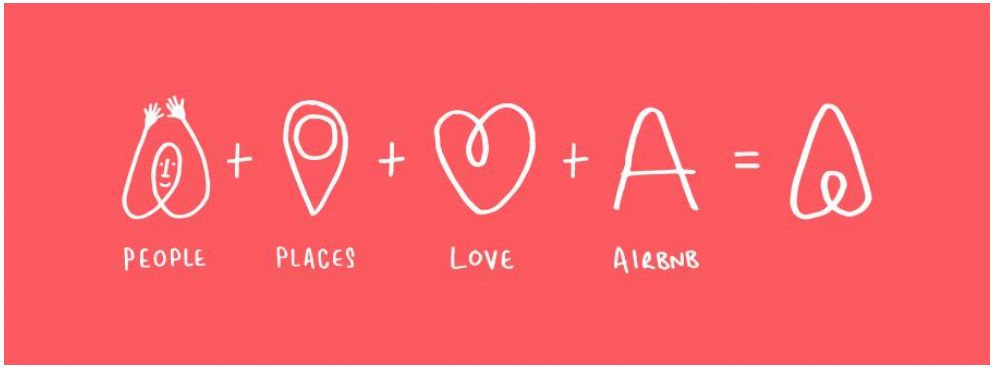
創辦人：Brian Chesky & Joe Gebbia

創辦年份：2007

最初的名稱：AirBed & Breakfast

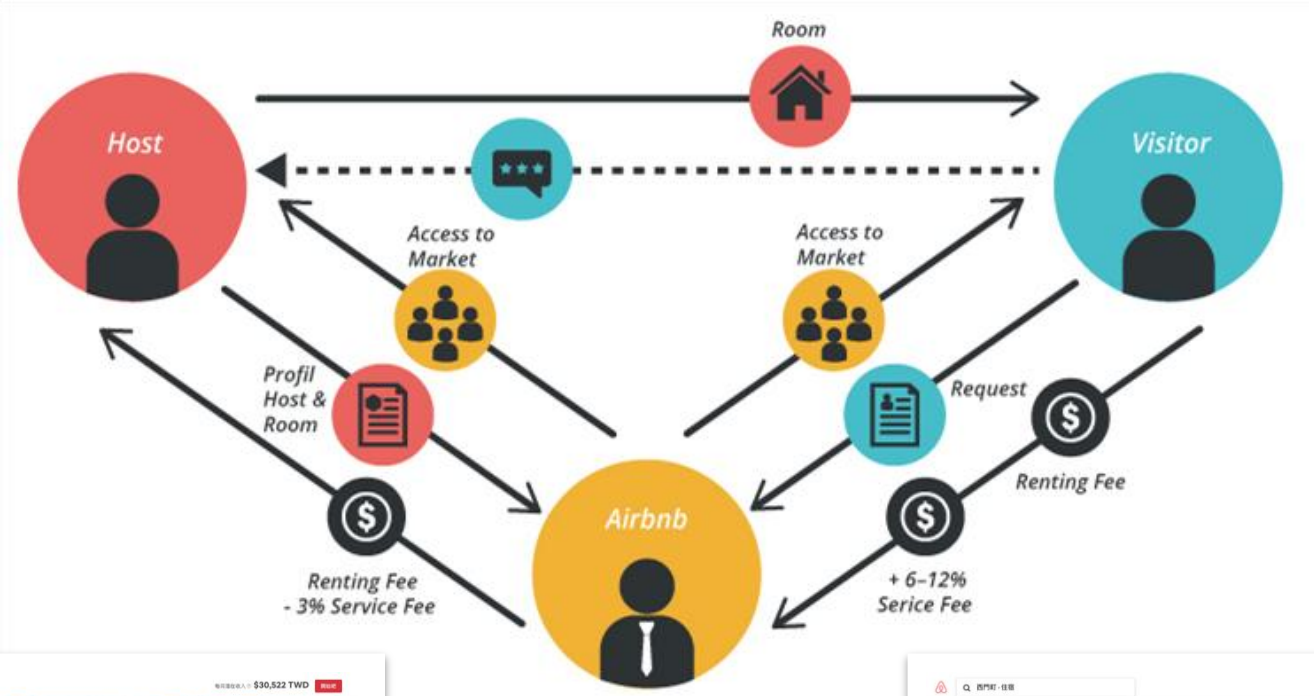
在191個國家、65,000個城市中
共有超過3,000,000筆房源

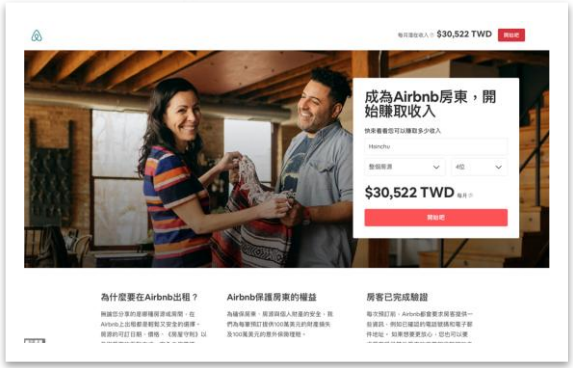
Airbnb是一個讓大眾出租住宿民宿的網站，提供短期出租房屋或房間的服務。讓旅行者可以通過網站或手機、發掘和預訂世界各地的各種獨特房源，為近年來共享經濟發展的代表之一。

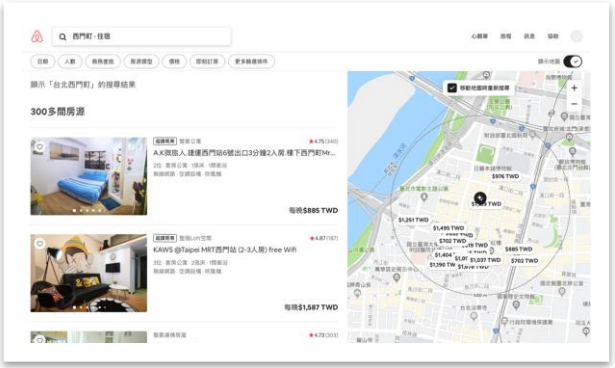


Airbnb 背景與專題主題介紹

Business Model







Airbnb 背景與專題主題介紹

Challenges & Project Object

Challenge

每個獨一無二的房間

如何訂出最適當的價格？

Hosts：收入最大化

Visitors：最划算

Airbnb：滿意度、訂房率提昇

Airbnb官方研究：Aerosolve: Machine learning for humans

- Airbnb如何提高留客率？星等不準，靠AI分析4百萬家民宿所有評論，才能找出值得推薦的好房東 | iThome
- Airbnb 數據團隊主管：把數據科學家的工作分成「三種方向」，才能發揮最大效益！ | TechOrange
- Airbnb Engineering & Data Science – Medium
- 【演講精華濃縮】Airbnb 資料科學部主管：企業如何以數據為核心打造精準決策？ | TechOrange
- Yahoo看見數位行銷力 — Airbnb怎麼用大數據說出一個百億估值的故事？
- Airbnb如何靠數據成功？3點揭密共享經濟下的數據顯學 | SmartM 新網路科技
- Airbnb 如何利用大數據幫使用者確定房租價格？ | TechNews 科技新報
- 出租率增四倍 Airbnb大數據策略揭祕 - 今周刊
- GitHub - airbnb/aerosolve: A machine learning package built for humans.
- The Secret Of Airbnbs Pricing Algorithm - IEEE Spectrum
- Jeff Feng: Head of Machine Learning and Analytics Product, Airbnb - YouTube
- Airbnb Pricing Tools | 2019 Updated Best Tools Report
- Here come the hackers: Airbnb's pricing algorithm | The FCT Blog

數據分析流程

Work Flow

Inside Airbnb
Adding data to the debate

airbnb

資料收集　　資料清洗　　機器學習

the A team

ABOUT DATA

原始資料來源及描述

原始資料來源及描述

Raw Data & Brief Description

the A team

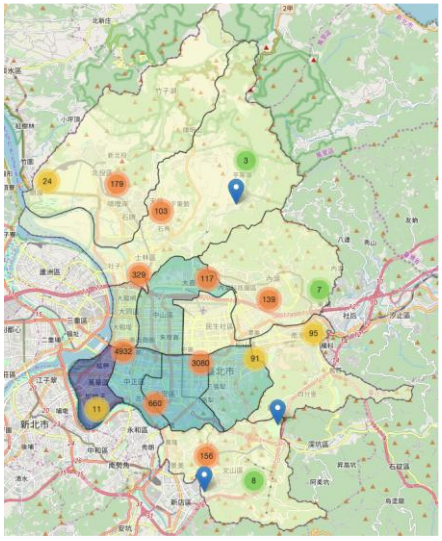
Inside Airbnb
Adding data to the debate

Listings

Reviews

	listings.csv 房源資料	reviews.csv 住客評論
原始欄位	106	6
資料筆數	142,628	334,018
資料期間	2018.Aug ~ 2019.Sep.	2011 ~ 2019.Sep.
資料類別	數值資料:43 類別資料:27 文字資料:30 時間資料:6	類別資料:3 文字資料:2 時間資料:1
內容	房源資料:65 房東資料:22 價格資料: 7 評論資料:22	房源id 房客id 評論日期 評論文字

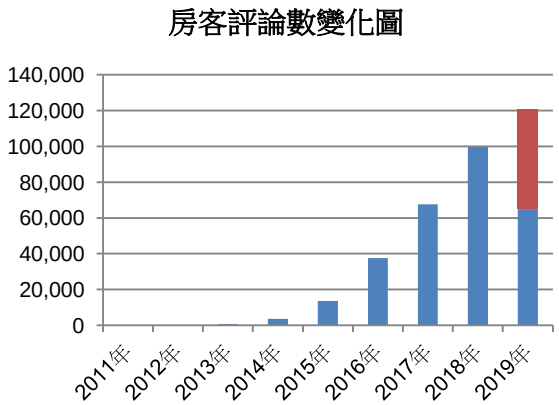
Airbnb Taipei

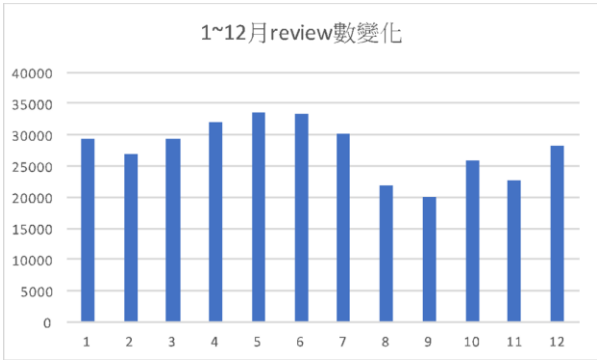


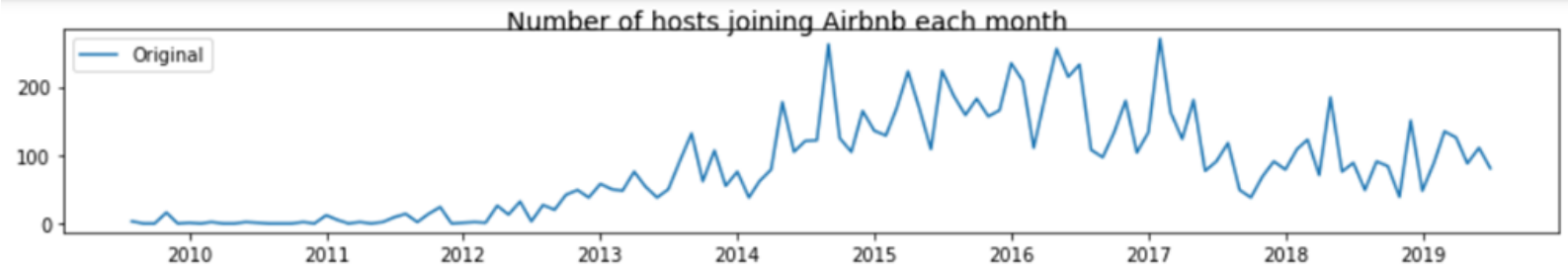
原始資料來源及描述

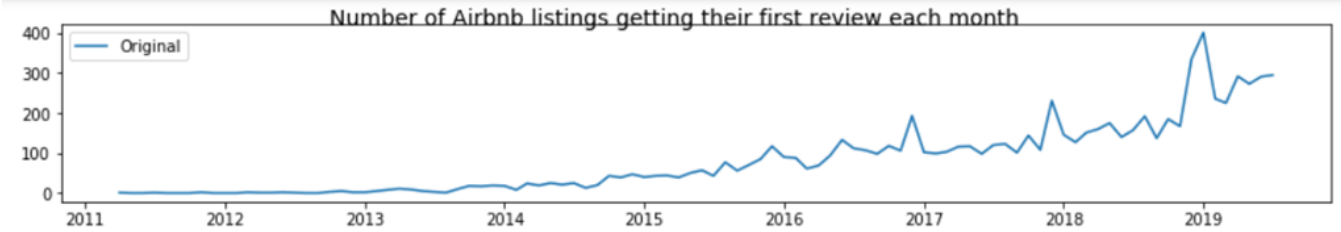
Raw Data & Brief Description

Reviews









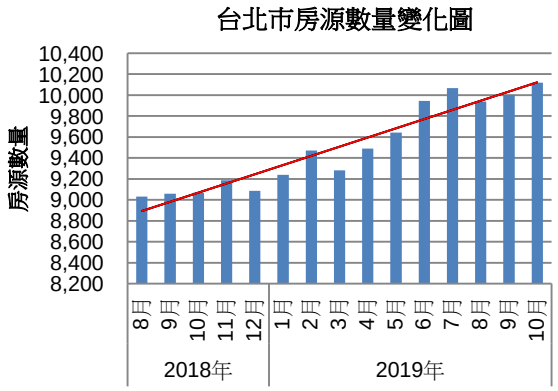
原始資料來源及描述

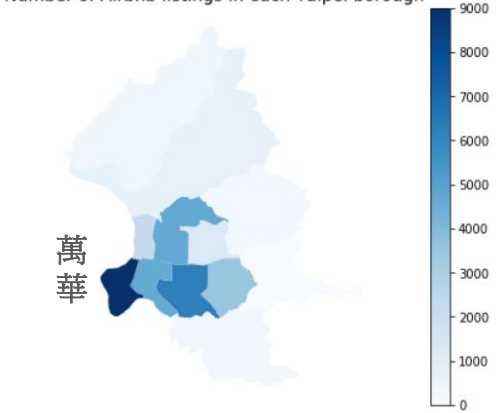
Raw Data & Brief Description

the A team

python

Listings

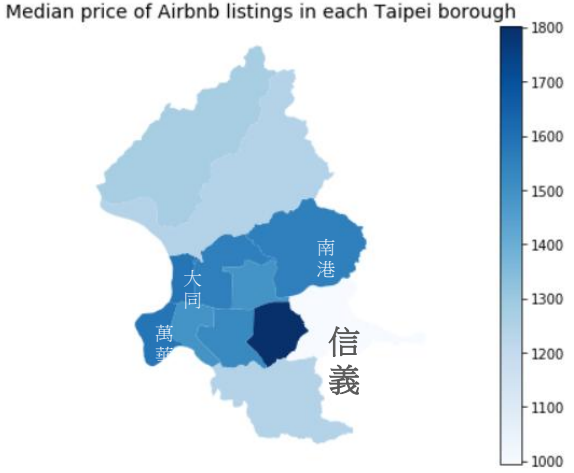


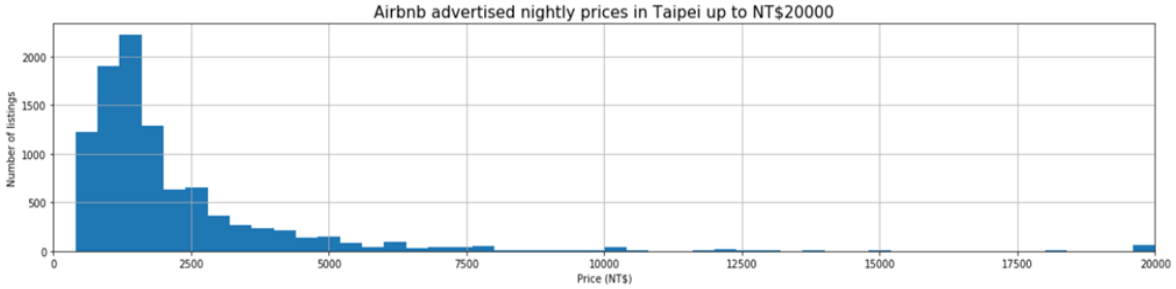


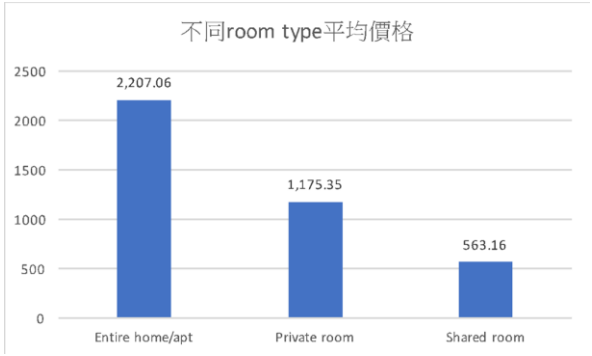
原始資料來源及描述

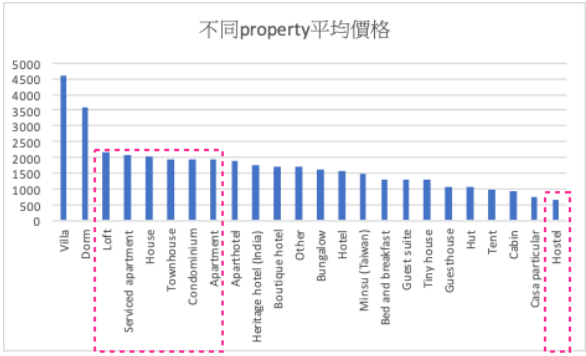
Raw Data & Brief Description

Listings









Questions & Challenges

1. 若要提供"最佳定價”，現有資料欄位是否都是必須的？
2. 現有資料內容包含空值！
3. 目前的實際定價(price)是否能作為模型訓練的”最佳定價”？
4. Reviews資料中的欄位是否有信息可以幫助定價？
5. 公開資料中並沒有經營數據（如：交易資料、住房率…）
6. 住客(visitor)信息只有id與用戶帳號名

目前的實際價格定對了嗎？

price
$1,555.00
$933.00
$1,555.00

```
In [12]: df_reviews.info()
         df_reviews.head(5)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 334016 entries, 0 to 334015
Data columns (total 6 columns):
listing_id       334016 non-null int64
id               334016 non-null int64
date             334016 non-null object
reviewer_id      334016 non-null int64
reviewer_name    334016 non-null object
comments         333524 non-null object
dtypes: int64(3), object(3)
memory usage: 15.3+ MB
```

Out[12]:

	listing_id	id	date	reviewer_id	reviewer_name	comments
0	68396	4743816	2013-05-23	3334370	Scott	Great location (very close to the metro stati...
1	68396	279164758	2018-06-20	183745170	Yamada	Very good place!!!\n
2	68398	2894700	2012-11-16	3534183	Laura	I enjoyed my stay very much! Spend approx 6 we...
3	68398	13416072	2014-05-27	12750484	Maggie	I ended up staying in Studio B. Apartment is c...
4	68398	22427468	2014-11-06	21131563	Vincent	Lisa est très efficace, le logement est foncti...

the A team

Data Cleaning

資料清洗

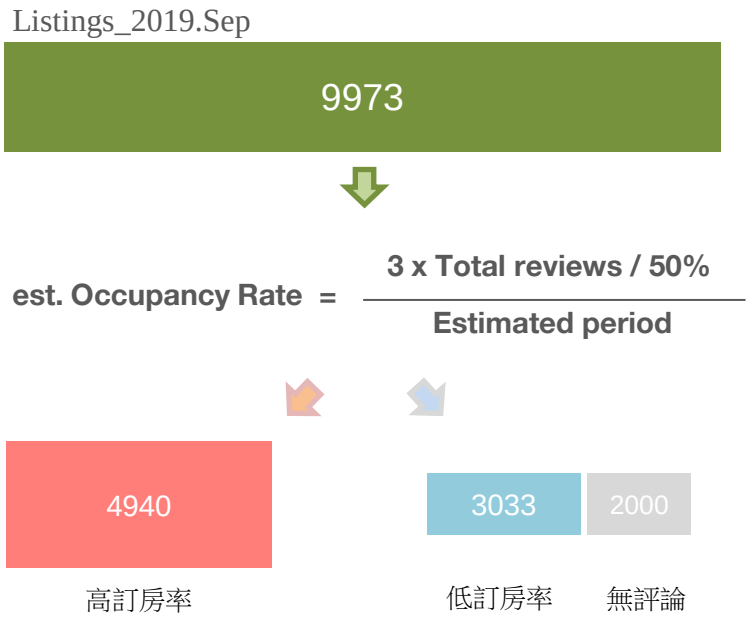
Listings

Reviews

定義最佳定價
"Good" Pricing?

價格定得好 >> 訂房率較好

Questions: 資料中未提供訂房率



使用2019.Sep 資料共9973筆，根據定義篩選訂房率後，
資料筆數為4940筆，以這些房源進行模型訓練

Total # bookings = # Total reviews / Review rate
estimated # bookings = # Total reviews / 50%
Occupancy = 3 x estimated # bookings

Reference:

- A high availability metric and filter of 60 days per year
- A frequently rented filter of 60 days per year
- A review rate of 50% for the number of guests making a booking who leave a review
- An average booking of 3 nights unless a higher minimum nights is configured for a listing.
- A maximum occupancy rate of 70% to ensure the occupancy model does not produce artifically high results based on the available data
- 2018年全台民宿的平均房客住用率為 20.09%，參考 https://reurl.cc/nVXgrX

Listings欄位資料清洗

Data Cleaning

資料清理

- 刪除不相關/重複
- 定義模糊
- 地理位置相關
- 刪除空值過高(95%>)
- 空值補平均值/中位數

資料轉換

- 屬性單一化 
- 價錢類轉為數值
- 時間類轉為天數(days)，再轉為類別
- host_listings_count由數值轉為類別

資料整合

- Amenities中的文字資料用pandas索引切片後選出Top 20設施轉為類別 
- 經緯度欄位合併用以後續計算

資料簡化

- 替換離峰值/異常值
- 連續欄位標準化

Listings

數值資料:43
類別資料:27
文字資料:30
時間資料:6

→

數值資料:43→22
類別資料:27→12
文字資料:30→4
時間資料:6→3

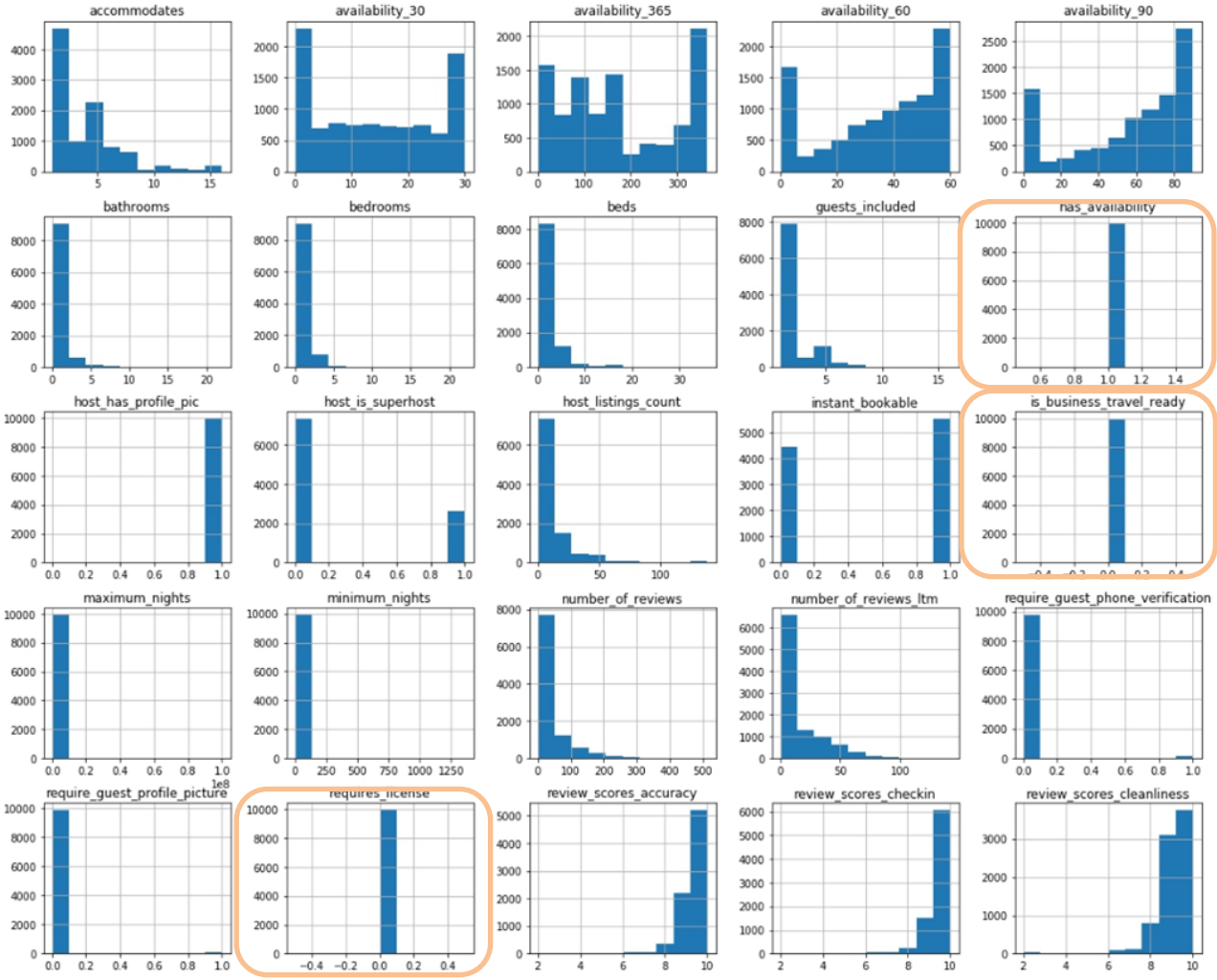
→

數值資料:22→25
類別資料:12→29
~~文字資料:30→0~~
~~時間資料:6→0~~

Python Code

Listings欄位資料清洗

Data Cleaning



Python Code

Listings欄位資料清洗

Data Cleaning

the A team

pandas matplotlib python

```
In [4]: df.amenities.head(3)
```

```
Out[4]: 0    {TV,"Cable TV",Internet,Wifi,"Air conditioning",Doorman,Elevator,"Family/kid friendly",Washer,Drye
        r,"Smoke detector","Fire extinguisher",Essentials,Shampoo,"Lock on bedroom door","24-hour check-in",Hanger
        s,"Hair dryer","Laptop friendly workspace"...
        1    {TV,"Cable TV",Wifi,"Air conditioning",Washer,Dryer,"Smoke detector","Fire extinguisher",Essentials,S
        hampoo,"Lock on bedroom door",Hangers,"Hair dryer","Laptop friendly workspace","Self check-in",Keypad,"Pri
        vate entrance","Hot water","Bed linens"...
        2    {TV,"Cable TV",Wifi,"Air conditioning",Kitchen,"Paid parking off premises","Smoking allowed",Elevato
        r,Washer,"Smoke detector","Carbon monoxide detector","Fire extinguisher",Essentials,Shampoo,"Lock on bedro
        om door",Hangers,"Hair dryer","Laptop fr...
        Name: amenities, dtype: object
```

```
In [7]: df.head(5)
```

Out[7]:

	Air conditioning	Wifi	Hair dryer	Shampoo	Essentials	Carbon monoxide detector	TV	Hangers	Washer	Laptop friendly workspace	Fire extinguisher	Smoke detector	Ho wa
0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	1.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0
4	1.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0

Python Code

Reviews文字資料清洗

Text Mining

BeautifuleSoup
NLTK

爬蟲

- Python BeautifulSoup
- 新增房客國別

(334,017)

判斷語種

- Python langdetect
- 英文佔最多數 39%

(287,572)

斷詞/詞性判斷

- Python nltk, takenize, pos_tag
- 斷出36種詞性 / 19,659字 /出現頻率 5,821K次
- 留下16種詞性並挑出cover 90%頻率, 計1,054字/2711K次
- 挑出459個相關字眼

(133,719)

文字向量計算

- TF
- IDF
- Word Vector

評論情感分析

- Python nltk.sentiment.vader

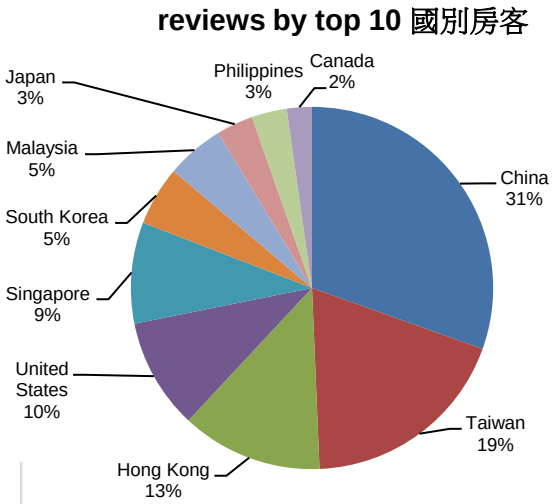
房源資料: 2
房客資料: 2
時間資料: 1
文字資料: 1

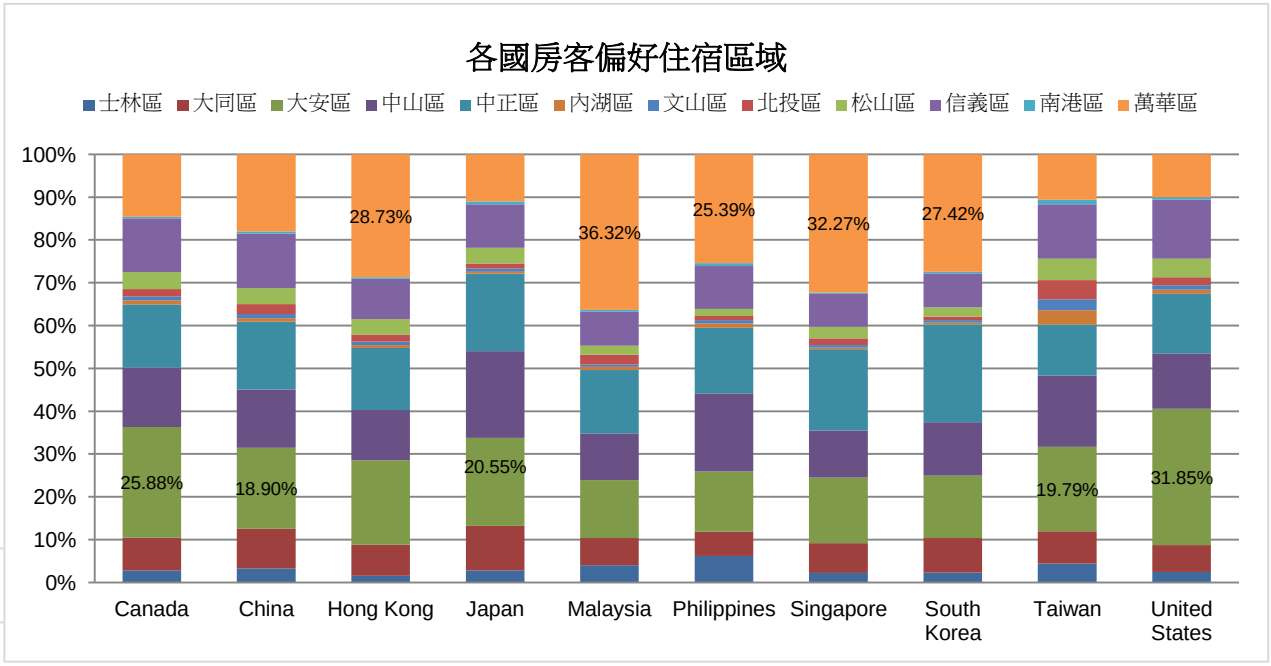
房源資料: 2
房客資料: 2→1
時間資料: 1
~~文字資料: 1→0~~
情感分析: 0→1
文字向量: 0→9

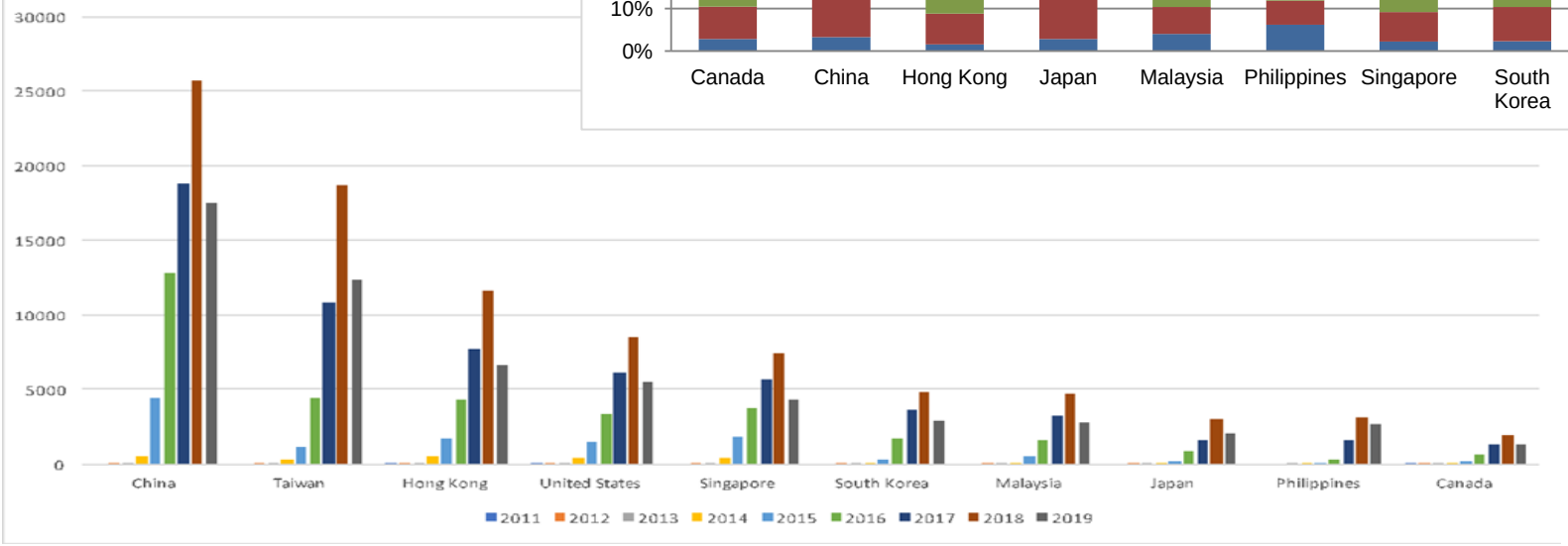
Reviews文字資料清洗

Foreigners in Taipei Airbnb

the A team







Reviews文字資料清洗

Text Mining

the A team

NEW TABLE

	listing_id	date	reviewer_id	country	senti_compound	txt_safety	txt_price	txt_host_feel	txt_host_real	txt_entert_food	txt_facility_sw	txt_faclity_hw	txt_cl
0	74643	2011-04-05	429351	United States	0.8576	0	0	1	1	2	2	2	
1	74643	2011-06-27	717646	Germany	0.9918	0	1	0	2	0	2	6	
2	178036	2011-07-30	274232	United States	0.9169	0	0	2	0	0	0	2	
3	74643	2011-09-15	894102	United States	0.9812	0	1	2	2	1	3	6	
4	178036	2011-10-02	1155635	Hong Kong	0.9450	0	0	0	1	0	0	0	

Reviews

```
RangeIndex: 126603 entries, 0 to 126602
Data columns (total 14 columns):
listing_id          126603 non-null int64
date                126603 non-null object
reviewer_id         126603 non-null int64
country             126603 non-null object
senti_compound      126603 non-null float64
txt_safety          126603 non-null int64
txt_price           126603 non-null int64
txt_host_feel       126603 non-null int64
txt_host_real       126603 non-null int64
txt_entert_food     126603 non-null int64
txt_facility_sw     126603 non-null int64
txt_faclity_hw      126603 non-null int64
txt_cleaness        126603 non-null int64
txt_loc_traffi      126603 non-null int64
dtypes: float64(1), int64(11), object(2)
memory usage: 13.5+ MB
```

	textmining_table.csv
欄位	14
資料筆數	126603
資料期間	2011 ~ 2019.Sep.
內容	房源id 房客id 房客國別 評論日期 情感分析分數 文字向量

Questions & Challenges

1. 該選用機器學習的哪種模型來進行定價建議？
2. 從review來的文字挖掘資料，可以發現什麼？

the A team

MODEL LEARNING

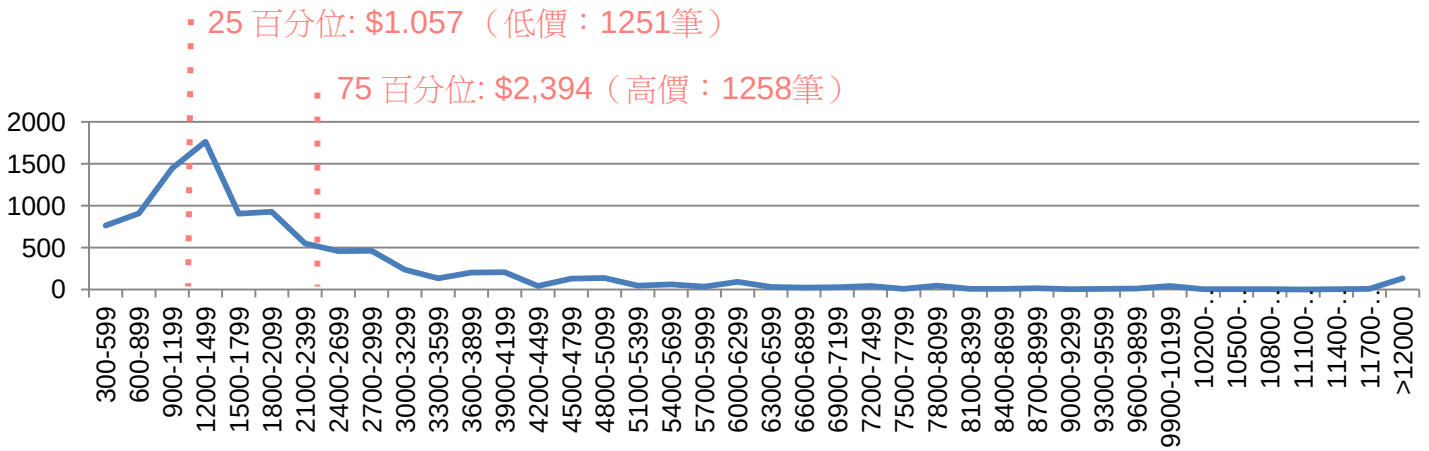
機器學習模型分析

文字分析

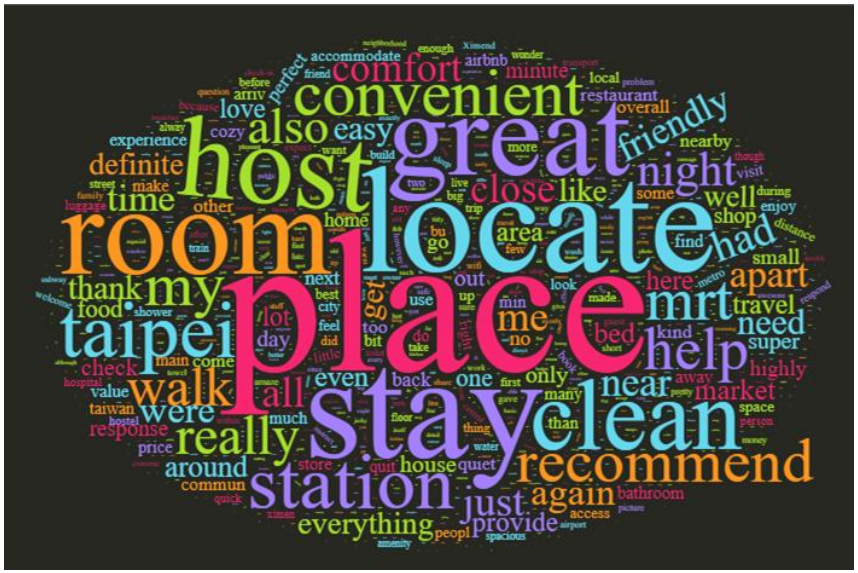
Text Mining

WordCloud python

Reviews



高價



低價

文字分析 區段差異

Text Mining

WordCloud python

高價

低價

高價	低價
大部分為家庭或團體旅行	大部分為個人旅行、背包客
空間大、設施、浴室、住幾人	空間不大、金錢、
安靜、便利	友善、體驗、捷運站

選擇機器學習模型
Model Selection_Prediction

the A team

python

預測目標	變數數量	資料對象
訂房價格	54	高訂房率的房間

Listings

預測

何謂準確：| 預測價格 - 實際價格 | / 實際價格 < 20% 的資料比例

MODEL	Multilayer Perception	SMOreg	Random Forest	Random Tree
準確度	22.67%	46.02%	50.78%	26.96%
overfitting	45.27%	0.04%	36.79%	64.53%

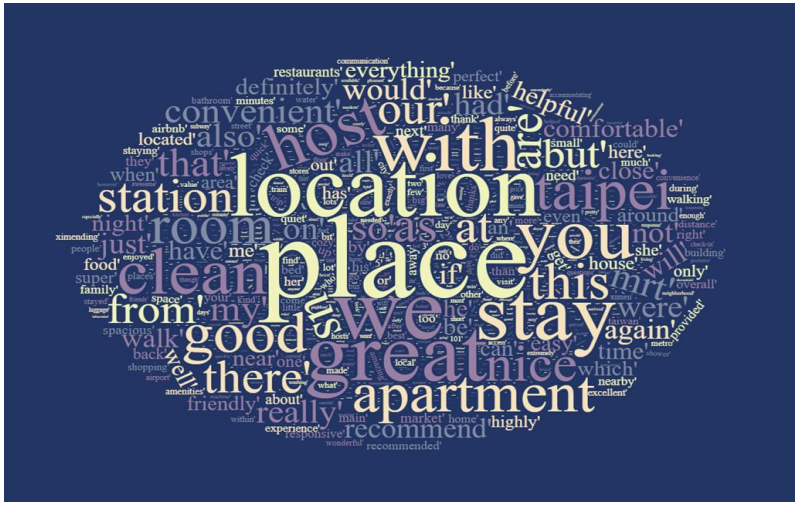
挑選出
10個變數

MODEL	Multilayer Perception	SMOreg	Random Forest	Random Tree
準確度	44.38%	45.81%	47.24%	41.32%
overfitting	2.59%	3.634%	30.01%	43.10%

運用Text Mining的發現

對「位置、距離」的重視

WordCloud python



“位置”是評論中，最常被提到的！

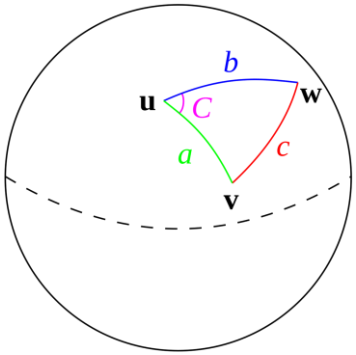
```
import csv
from math import radians, cos, sin, asin, sqrt

def haversine(lon1, lat1, lon2, lat2): # 經度1，緯度1，經度2，緯度2
    # 轉換成弧度
    lon1, lat1, lon2, lat2 = map(radians, [lon1, lat1, lon2, lat2])

    # haversine公式
    dlon = lon2 - lon1
    dlat = lat2 - lat1
    a = sin(dlat/2)**2 + cos(lat1) * cos(lat2) * sin(dlon/2)**2
    c = 2 * asin(sqrt(a))
    r = 6371 # 地球平均半徑
    return c * r * 1000 #單位公尺

#longitude 經度
#latitude 緯度
with open('listings-2 lalo.csv', newline='',encoding='UTF-8') as cf:
    reader = csv.DictReader(cf)
    idList= []
    lon1List = []
    lat1List = []
    aList = []
    lon2 = float(input('lon:'))
    lat2 = float(input('lat:'))
    p  = input('location:')
    with open(p+'.csv', 'w', newline='',encoding='utf8') as csvfile:
        writer = csv.writer(csvfile)
        writer.writerow(['id',p])
        for row in reader:
    #         print(row['id'],row['longitude'],row['latitude'])
            idList.append(row['id'])
            lon1List.append(float(row['longitude']))
            lat1List.append(float(row['latitude']))

        for i,(lon,lat) in enumerate(zip(lon1List,lat1List)):
            a = round(haversine(lon,lat,lon2,lat2), 2)
    #         print(lat,lon,lat2,lon2)
            aList.append({
                'id':idList[i],
                p:a
            })
            aL = aList[i]
            writer.writerow([aL['id'],aL[p]])
            print(aL['id'],aL[p])
```



The law of Haversines

[計算] 各房間與最近的捷運站之間的距離。
→ 增加為新的變數X

Reviews

選擇機器學習模型

Model Selection_Prediction

the A team

python

Listings

挑選出
10個變數

何謂準確：| 預測價格 - 實際價格 | / 實際價格 < 20% 的資料比例

MODEL	Multilayer Perception	SMOreg	Random Forest	Random Tree
準確度	44.38%	45.81%	47.24%	41.32%
overfitting	2.59%	3.63%	30.01%	43.10%

加入「距離最近捷運站的距離」

MODEL	Multilayer Perception	SMOreg	Random Forest	Random Tree
準確度	49.63%	46.22%	47.17%	38.39%
overfitting	3.36%	4.27%	39.97%	61.61%

小結與後續分析

Wrap Up & Future

一、我們的系統可提供房東合理的定價建議。

二、從文字雲的分析結果我們發現：

1.房客關注房源地點：
　　我們加入房源與大眾運輸系統的距離做為新的變數進行測試，結果對房價預測的準確率有明顯提升。

2. 高訂房率之高價房源與低價房源有所差別：
- 兩種價位，兩組不同的客群屬性，兩組不同的關注條件。
- 我們的系統可依房東的訂價高低，為房東提供不同的經營建議。

三、若能有更多房客的資訊或實際訂房率，相信我們的準確率會更加提升。

組員介紹
Team Members

陳國誠	李柏勳	徐旭洋	温子霈
背景領域：醫療、化妝品行銷 專題工作：組長 / 資料分析 / 工作與進度分配	背景領域：面板廠管理 專題工作: 資料預處理、資料視覺化	背景領域：行銷、企劃、業務、平面設計、網頁/前端設計、APP產品規劃...。 專題工作：資料分析/海報/網頁設計製作。	背景領域：日文、資管、工管 專題工作: 數據工程、資料視覺化
范瑞紋	**陳詠青**	**許呈安**	**謝昀叡**
背景領域：中文/研究倫理 教學 專題工作：摘要撰寫 /text mining初步選詞 / 系統應用介面	背景領域: 網路通訊廠 sales/logistic 專題工作: 資料分析	背景領域: 供應鏈管理 專題工作: 資料預處理與機器學習	背景領域: 半導體、IC 產業 專題工作: 資料清洗整理
李君明			
背景領域: 半導體產業、CISCO、FAB無人化 專題工作: 資料整理			