Airbnb

2019 Airbnb Analysis in New York City

How different factors influence the demand of several properties on Airbnb

February, 2023

The data from this analysis is from Kaggle New York City Airbnb Open Data.

Problems Definition

此資料分析主要希望找出：當顧客在選擇 Airbnb 時，

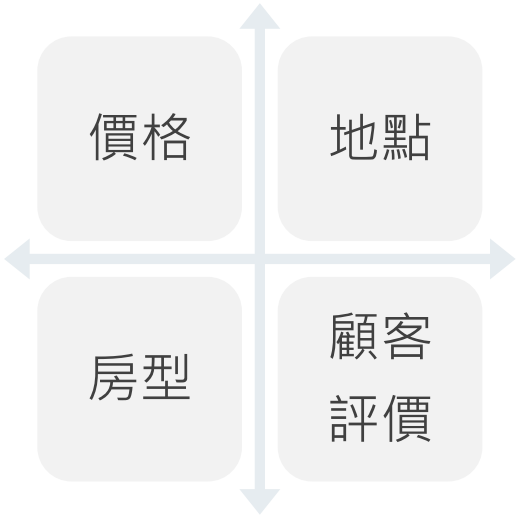
哪些因素對於他們的選擇影響最大，以及紐約市的 Airbnb 業者該以何種商業決策應對

定義問題

資料及內所有的欄位名稱			
id	neighbourhood_group	room_type	last_review
name	neighbourhood	price	reviews_per_month
host_id	latitude	minimum_nights	calculated_host_listings_count
host_name	longitude	number_of_reviews	availability_365

根據上表的欄位名稱和數據，此數據分析將從以下問題最為切入研究之方向：

- 哪一個區域最熱門
- 哪一種房型最受歡迎
- 不同的區域地點是否會影響房價和整體需求量
- **當顧客在挑選 Airbnb 時，什麼因素是最重要的（價格、地點、房型、顧客評價）**



Exploratory the Data

資料之間**沒有明顯的相關性**，但 Airbnb 的**價格與需求**

和其座落的**地點以及房型**有可能具有相關性

資料集 EDA

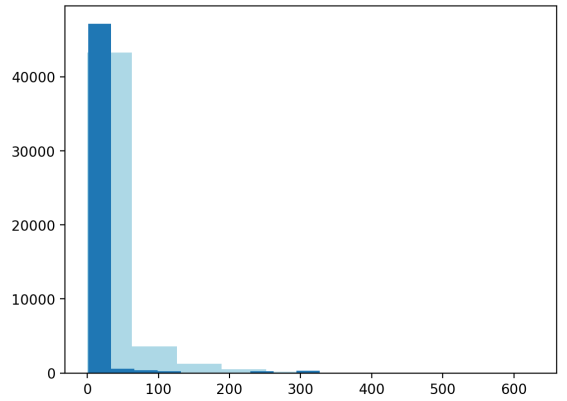
資料集之描述性統計（已刪除不需要的資料）

* 不需要的資料：id, host_id, name, ost_name

	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

結果

- price, minimum_nights 兩者有明顯的異常值
- number_of_reviews, calculated_host_listings_count 兩者皆高度右偏態
- 因此要對上述幾個有問題的資料做整理



資料集 EDA

確認資料是否有空值

```
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
last_review                     10052
reviews_per_month               10052
calculated_host_listings_count      0
availability_365                    0
dtype: int64
```

結果

- Last_review, reviews_per_month 兩者含有大量空值
- 因此要對他們進行資料整理（因不需要這兩個資料，因此直接刪掉以上兩個欄位資料）

資料清理

Step 1 - 移除離群值

	z_price	z_minimum_nights
count	48197.000000	48197.000000
mean	0.329868	0.315125
std	0.308686	0.303088
min	0.001163	0.001461
25%	0.155232	0.196484
50%	0.300974	0.245240
75%	0.406912	0.293996
max	2.945135	2.972649

利用 Z-score 來找出離群，並刪除 Z-score 絕對值大於 3 的數值

Step 2 - 將資料轉換為分類資料

minimum_nights_group	host_listing_group
one_night	more_listing
one_night	two_listing
three_night	one_listing
one_night	one_listing
more_night	one_listing

Step 3 - 刪除不要的欄位

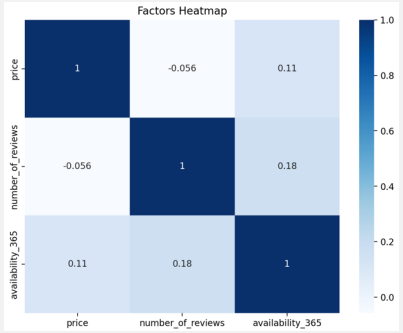
最終的資料集

	latitude	longitude	price	number_of_reviews	availability_365
count	48197.000000	48197.000000	48197.000000	48197.000000	48197.000000
mean	40.728856	-73.951896	138.749258	23.479034	111.656618
std	0.054603	0.046209	107.591518	44.735633	131.065455
min	40.499790	-74.244420	0.000000	0.000000	0.000000
25%	40.689930	-73.982830	69.000000	1.000000	0.000000
50%	40.722890	-73.955480	105.000000	5.000000	43.000000
75%	40.763130	-73.935900	175.000000	24.000000	223.000000
max	40.913060	-73.712990	860.000000	629.000000	365.000000

資料視覺化

相關性分析

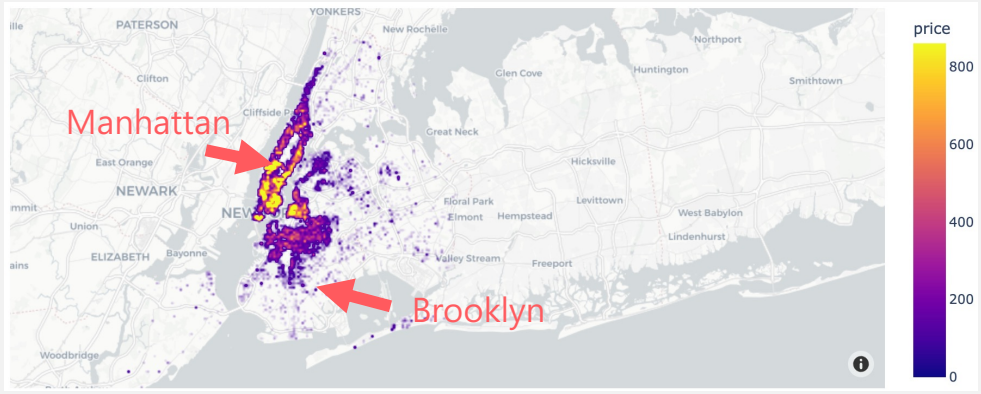
熱力圖與相關性矩陣 - 資料之間**沒有**明顯的相關性



	price	number_of_reviews	availability_365
price	1.000000	-0.056315	0.113493
number_of_reviews	-0.056315	1.000000	0.177297
availability_365	0.113493	0.177297	1.000000

從左邊兩張圖可以看出資料之間是幾乎沒有任何直接的相關性的

畫出 Price Density Map ，發現 Airbnb 在 Manhattan 和 Brooklyn 北邊較為集中，且租金也較高



其他可能會受區域影響的因子

	price			availability_365			number_of_reviews		
	count	mean	median	count	mean	median	count	mean	median
neighbourhood_group									
Bronx	1077	84.619313	65.0	1077	165.313835	147.0	1077	26.294336	9.0
Brooklyn	19915	116.504745	90.0	19915	99.557821	27.0	19915	24.312327	6.0
Manhattan	21201	174.996274	149.0	21201	110.187350	34.0	21201	21.269516	4.0
Queens	5635	94.125998	75.0	5635	143.973026	97.0	5635	27.798403	7.0
Staten Island	369	96.138211	75.0	369	198.934959	216.0	369	31.276423	12.0

數量最多： Manhattan

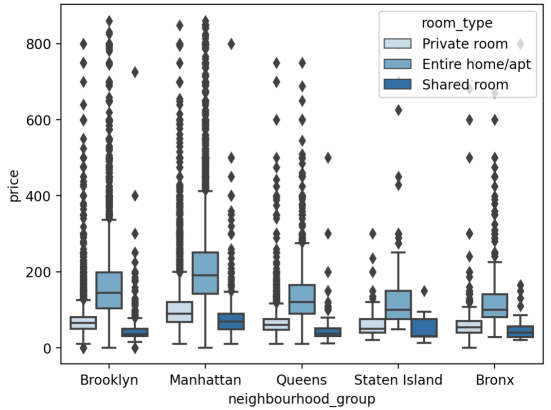
租金最高： Manhattan

需求最高： Brooklyn

資料視覺化

以 Boxplot 觀察價格和需求在不同類別之間的不同

不同區域和房型的價格差異



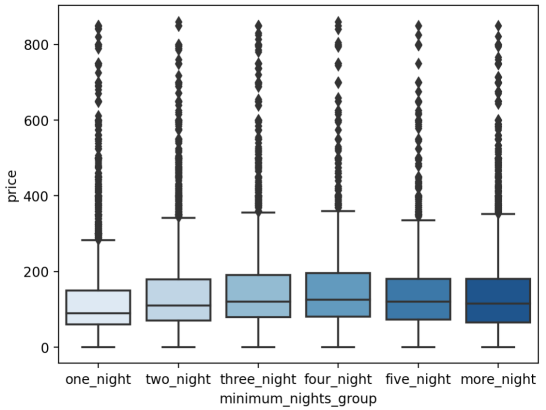
Price by Room Type

Entire Home/apt > Private Room > Shared Room

Price by Location

Manhattan > Brooklyn > Others

不同 minimum nights 的價格差異



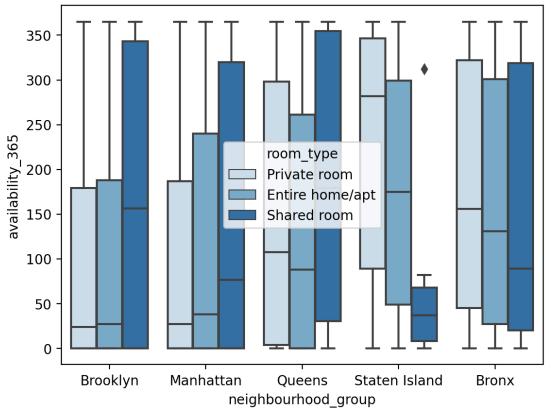
Cheapest

one_night

Most expensive

four/five_night

不同區域和房型的需求差異



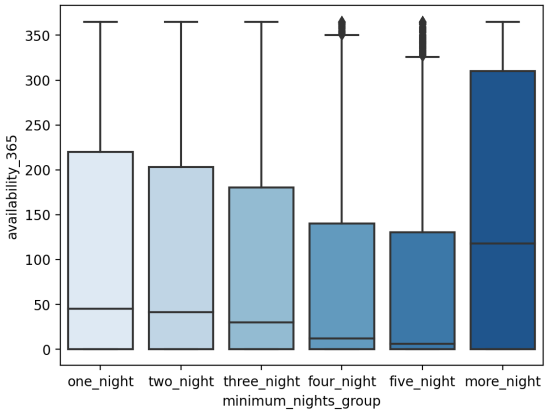
Demand by Room Type

Varies by locations

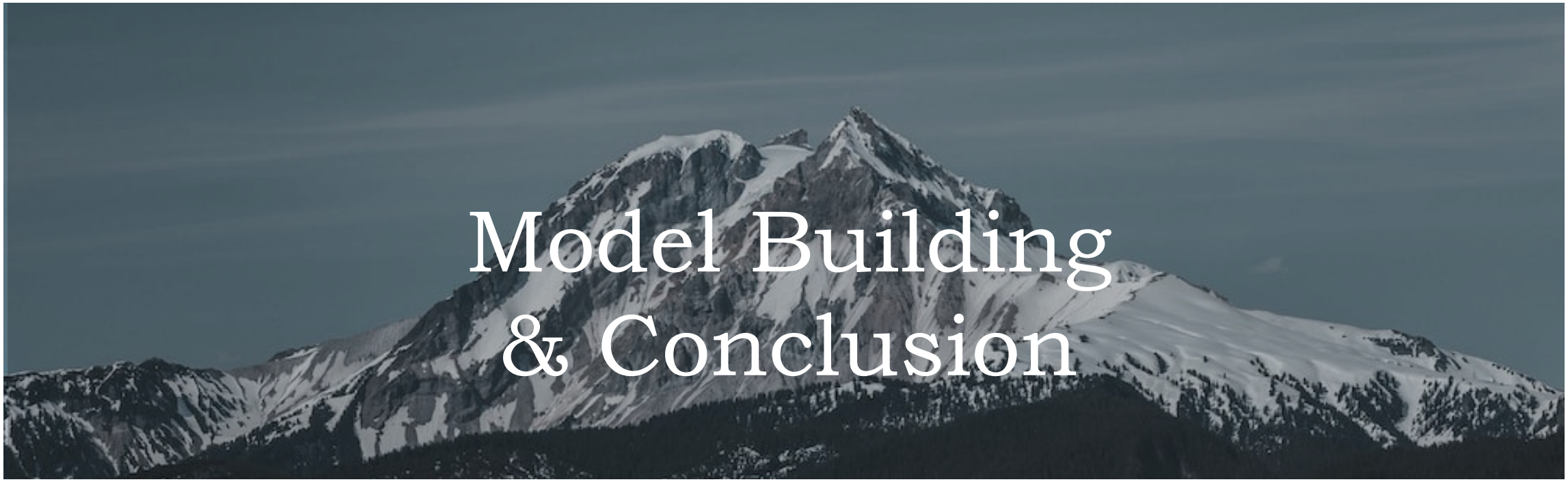
Demand by Location

- Brooklyn entire home have the highest demand, followed by Manhattan's entire home
- Staten Island properties have the lowest demand

不同 minimum nights 的需求差異



Five nights minimum are the most popular overall



Model Building & Conclusion

此研究所建立之模型主要是希望提供 Airbnb 業者如若想要對其現有或在規劃中的住宅做出需求的預測，一個較好的**回歸模型（隨機森林）**。並且得知**價格**和**評論數量**是業者需要特別關注的兩個重點。

Model 1 – Multiple Linear Regression (Statmodels / Sklearn)

將 Categorial Variables 轉換為 Numerical Variables

	price	number_of_reviews	availability_365	neighbor_frq	roomtype_frq	min_nights_frq	host_listing_frq
0	149	9	365	0.413200	0.459780	0.260825	0.203000
1	225	45	355	0.439882	0.516422	0.260825	0.136772
2	150	0	365	0.439882	0.459780	0.165010	0.660228
3	89	270	194	0.413200	0.516422	0.260825	0.660228
4	80	9	0	0.439882	0.516422	0.201942	0.660228
...	...	...	...	...	...	...	...
48890	70	0	9	0.413200	0.459780	0.241405	0.136772
48891	40	0	36	0.413200	0.459780	0.068096	0.136772
48892	115	0	27	0.439882	0.516422	0.201942	0.660228
48893	55	0	2	0.439882	0.023798	0.260825	0.203000
48894	90	0	23	0.439882	0.459780	0.201942	0.660228

- 因為 Categorial Variables 會讓建模變得很不好做，因此在建模前先將他們轉換為純數字的 Numerical Variables
- 而此資料分析當中，使用 **Frequency Encoding** 來做資料轉變

將 data 分為 training 和 testing 兩種類別

```
train_x.shape (33737, 6)
train_y.shape (33737,)
test_x.shape (14460, 6)
test_y.shape (14460,)
```

← 以 **70/30** 做分組

Model 1 – Multiple Linear Regression (Statmodels / Sklearn)

Statmodels

Dep. Variable:	availability_365	**R-squared:**	0.180
Model:	OLS	**Adj. R-squared:**	0.180
Method:	Least Squares	**F-statistic:**	1768.
Date:	Wed, 22 Feb 2023	**Prob (F-statistic):**	0.00
Time:	20:33:15	**Log-Likelihood:**	-2.9859e+05
No. Observations:	48197	**AIC:**	5.972e+05
Df Residuals:	48190	**BIC:**	5.973e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	235.1746	4.450	52.849	0.000	226.453	243.896
price	0.2253	0.005	42.209	0.000	0.215	0.236
number_of_reviews	0.4403	0.012	36.098	0.000	0.416	0.464
neighbor_frq	-127.4312	4.683	-27.213	0.000	-136.609	-118.253
roomtype_frq	-83.3523	7.479	-11.146	0.000	-98.010	-68.694
min_nights_frq	74.4588	8.745	8.515	0.000	57.319	91.599
host_listing_frq	-185.7584	2.400	-77.397	0.000	-190.463	-181.054

- 在 P-value 小於 0.05 的情況下，R-Squared 為 0.180 是明顯很小的，表示**資料之間不存在線性關係**
- price 和 number_of_review 都是正的，表示這兩項**數值越高，其對應的 Airbnb 需求則越低**

Sklearn

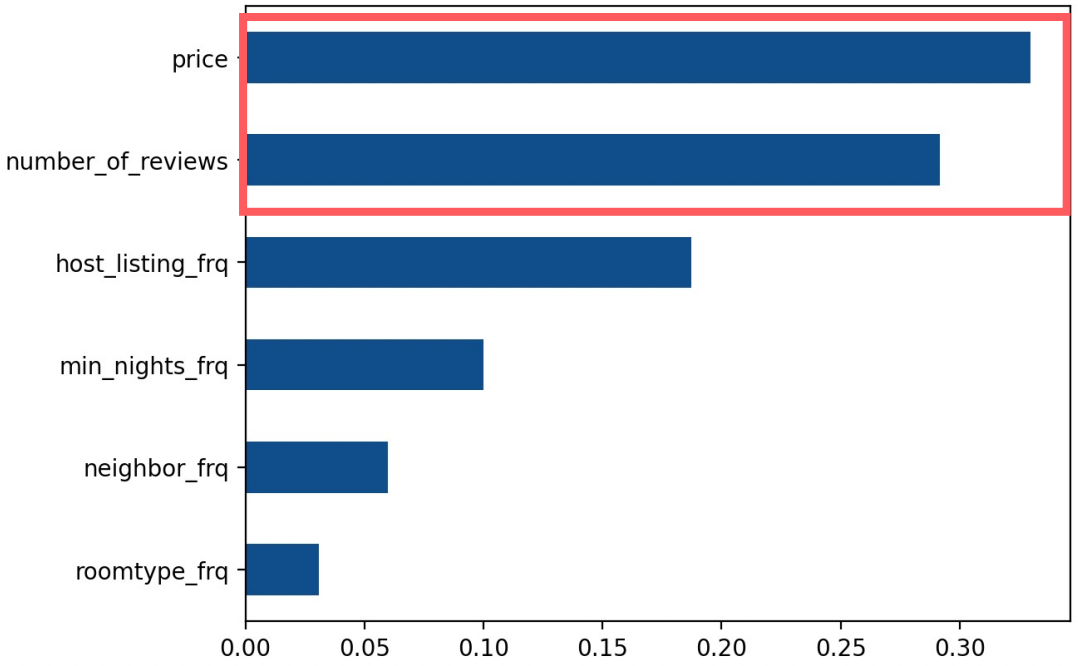
	Actual	Predicted
3098	0	58.731739
19134	0	52.999341
13939	0	49.920533
15189	0	70.967355
858	311	68.493441

- 在 Sklearn 底下，得出的結果為 **0.180498258515565**，和使用 Statmodels 的結果相似，再次確立了此資料及資料之間**沒有線性關係**。

Model 2 – Random Forest Regressor

使用 Random Forest Regressor 得出的結果為 **0.7884947345959598**，遠高於用線性回歸時得出的數值，表示此資料及在這格模型下的顯著性較高，以**這個模型來預測和解釋資料是比較合適的**。

Conclusion



左圖表示各種因子對於 Airbnb 需求的影響

可以看出**價格**和**評論的數量**是影響最大的，因此 Airbnb 的經營者應著重於對這兩點進行策略發想，來提高市場對於其 Airbnb 的熱度及需求