

# YOUTBIKE 2.0

## 借車人數預測

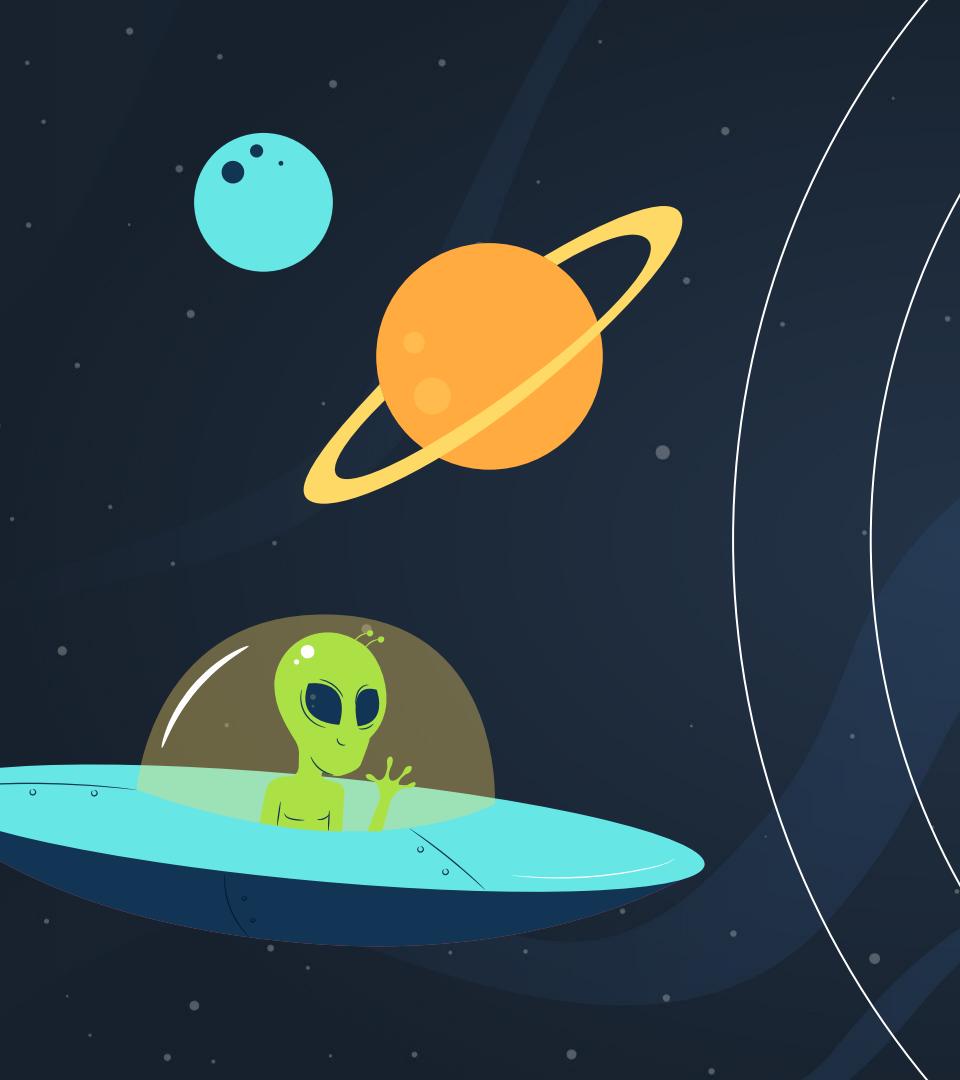
組員：

經濟五 張寬新  
經濟五 陳又加  
經濟五 蔡淳如  
經濟四 吳昕晏  
經濟四 錢紫翎

• Group : 阿北哩底隊

# AGENDA





01

# FRAME THE PROBLEM

# FRAME THE PROBLEM

## □ 問題

近期時常在學校的交流版上看到Youbike站點無車可借或是無位可還的情形發生，甚至有同學要從社科院前往大門，沿路上經過的9個站點都無車可借。此一情形違背了當初Youbike2.0設置讓同學隨時、隨地都有腳踏車可以租借使用的初衷，也更凸顯了資源分配的重要性。



# FRAME THE PROBLEM

## □ 目標

為了幫助台大校園內的 Youbike 2.0 系統能夠更有效率的被使用，我們希望能透過預測各站點一段時間後的使用情形，協助 Youbike 的營運方能對車輛做最適當的調度。



# FRAME THE PROBLEM

## □ 分析目標

我們認為影響**Youbike**站點使用情形的因素是非常複雜的，包括當時站內剩餘的車輛數、是否為上課日、當時的氣溫及降雨情形等都可能影響使用者的借車意願。另外，台大校內腳踏車拖吊系統的存在也是整個校園腳踏車生態中非常重要的因素，該站點所在地點或是附近地點是否有違規車輛被拖吊也可能影響 **Youbike**的借出情形



02

## DATA ACQUISITION

# DATA ACQUISITION

## 腳踏車相關資料

- (1) 自政府資料開放平臺上的**YouBike2.0**臺北市公共自行車即時資訊網站爬取**Youbike 2.0** 臺大專區中各站的即時站況資料，每15分鐘抓取一次
  
- (1) 於學校自行車線上管理系統網頁上爬取即時拖吊至水源腳踏車拖吊場的腳踏車輛資料，每15分鐘抓取一次
  
- (1) 向學校事務處申請取得**2019年至2021年10月底**的自行車拖吊資料

The screenshot shows the Data.gov.tw website interface. At the top, there's a navigation bar with links for '網站導覽' (Website Navigation), social media icons for Facebook and Twitter, and a search bar. Below the navigation, a breadcrumb trail shows the user is at the '資料集' (Dataset) level, specifically for the 'YouBike2.0臺北市公共自行車即時資訊' (YouBike2.0 Taipei Public Bike Real-time Information) dataset. The main content area displays the dataset title, a brief description in Chinese, and a rating section with a 4.8 star average from 15 reviews. Below this, there's a detailed view of a specific dataset entry, showing metrics like '瀏覽次數: 5161' (Viewed 5161 times), '下載次數: 1781' (Downloaded 1781 times), and '意見數: 1' (Comment count: 1). A note below explains the schema for the data columns. At the bottom of the page, there are download links for '資料資源下載網址' (Resource download address) and 'JSON' (JSON file), along with a link to the full dataset page.

The screenshot shows the homepage of the National Taiwan University (NTU) bicycle management system. It features the NTU logo and name, and the text '自行車線上管理系統' (Bicycle online management system) in both Chinese and English. Below this, there's a '最新公告' (Latest Announcement) section with three entries. The first entry is dated December 1, 2021, and mentions the announcement of bicycle parking rules and notices. The second entry is dated September 18, 2019, and discusses the identification and handling of abandoned bicycles. The third entry is dated August 13, 2021, and is about the auction of seized bicycles. There are also links for '自行車車證領取&注意事項(Announcement!!!)&大一女宿舍周邊自行車停放區分區淨空案11月起恢復執行' (Bicycle license application & notice matters (Announcement!!!) & the restoration of bicycle parking zones around female dormitories starting in November) and '廢棄自行車之認定與處理' (Identification and handling of abandoned bicycles).

# DATA ACQUISITION

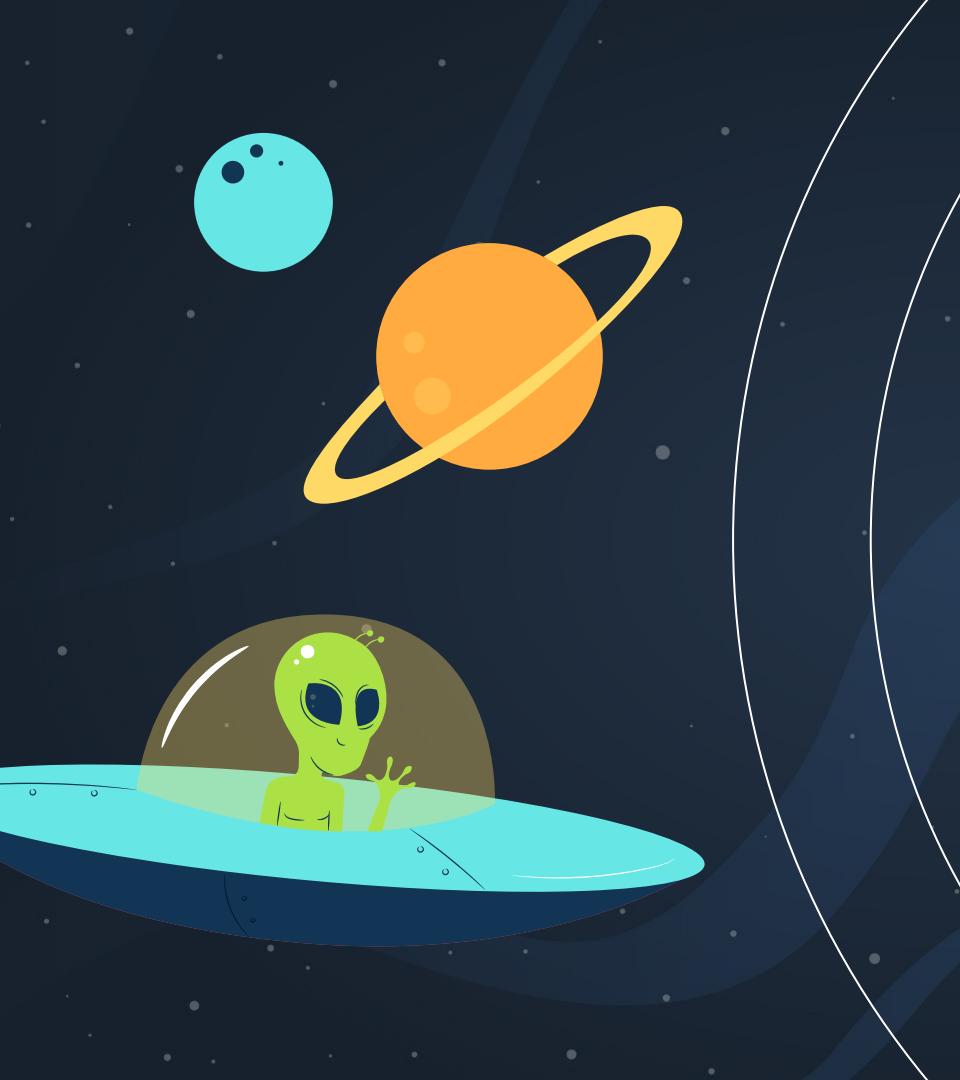
## 天氣資料

從交通部中央氣象局網站取得台灣大學測站觀測資料，包含氣溫、降雨等資訊，每小時紀錄一筆資料。



The screenshot shows the Central Weather Bureau's website interface. At the top, there is a navigation bar with links to Home Page, EN, Website Guide, Feedback, Common Questions, About the Bureau, and language selection (Small, Medium, Large). Below the navigation bar, there is a breadcrumb trail: 警特報 > 天氣 > 目前天氣 > 縣市測站列表 > 臺灣大學. The main content area is titled "臺灣大學測站觀測資料" (Observation Data of Taiwan University Station) and includes tabs for "過去24小時資料" (Past 24-hour Data), "過去24小時變化圖" (Past 24-hour Change Map), and "測站地圖位置" (Station Map Location). The data is presented in a table with the following columns: 觀測時間 (Observation Time), 溫度 (Temperature °C), 天氣 (Weather), 風向 (Wind Direction), 風力 (Wind Force), 陣風 (Gust), 能見度 (Visibility), 相對濕度 (%) (Relative Humidity %), 海平面氣壓 (Barometric Pressure), 當日累積雨量 (Daily Accumulated Rainfall), and 日照時數 (Solar Radiation Hours). The table contains 8 rows of data corresponding to observations from 12/28 16:00 to 12/28 17:10.

觀測時間	溫度 °C	天氣	風向	風力 (級)	陣風 (級)	能見度 (公里)	相對濕度 (%)	海平面氣壓 (百帕)	當日累積 雨量(毫米)	日照時數
12/28 17:10	17.5	多雲	東北東	2	-	-	86	-	0.0	-
12/28 17:00	17.6	多雲	東北	2	3	-	86	-	0.0	-
12/28 16:50	17.7	多雲	東北	1	-	-	85	-	0.0	-
12/28 16:40	17.8	多雲	東北	2	-	-	84	-	0.0	-
12/28 16:30	17.9	多雲	北北東	1	-	-	84	-	0.0	-
12/28 16:20	18.0	多雲	東北	1	-	-	83	-	0.0	-
12/28 16:10	18.2	多雲	東北東	1	-	-	82	-	0.0	-
12/28 16:00	18.3	多雲	東北東	2	4	-	82	-	0.0	-

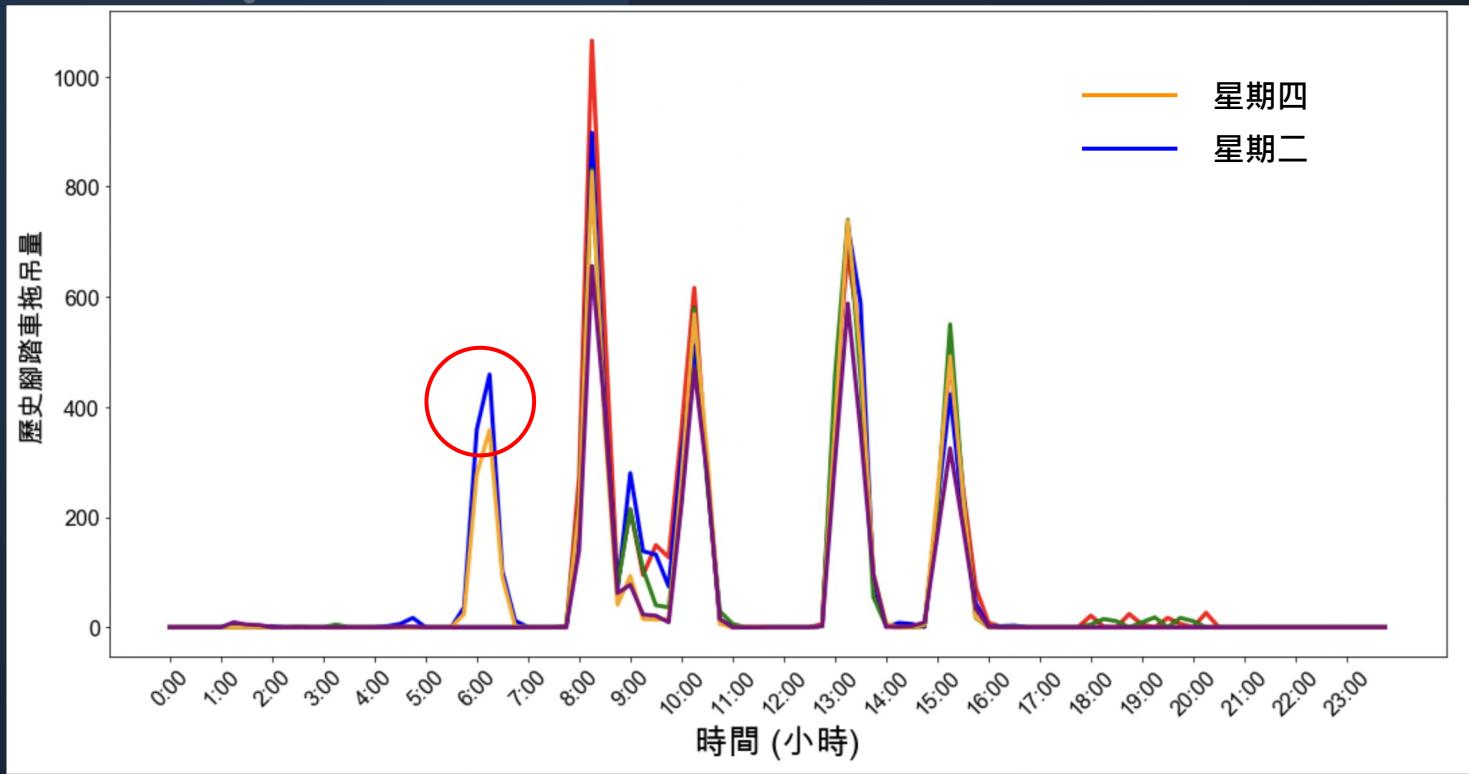


03

## EXPLORE THE DATA

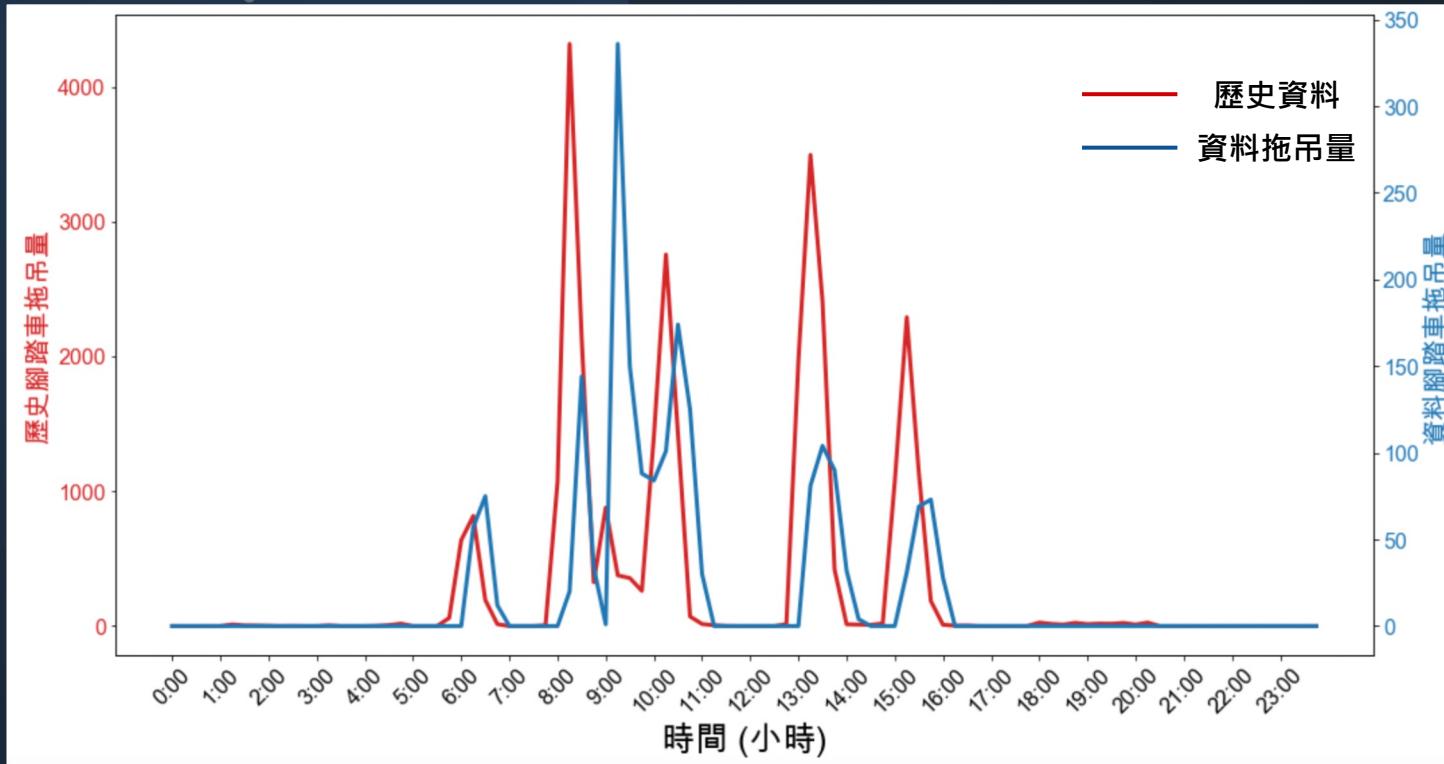
# EXPLORE THE DATA

- 歷史拖吊量(依照星期幾做區分)



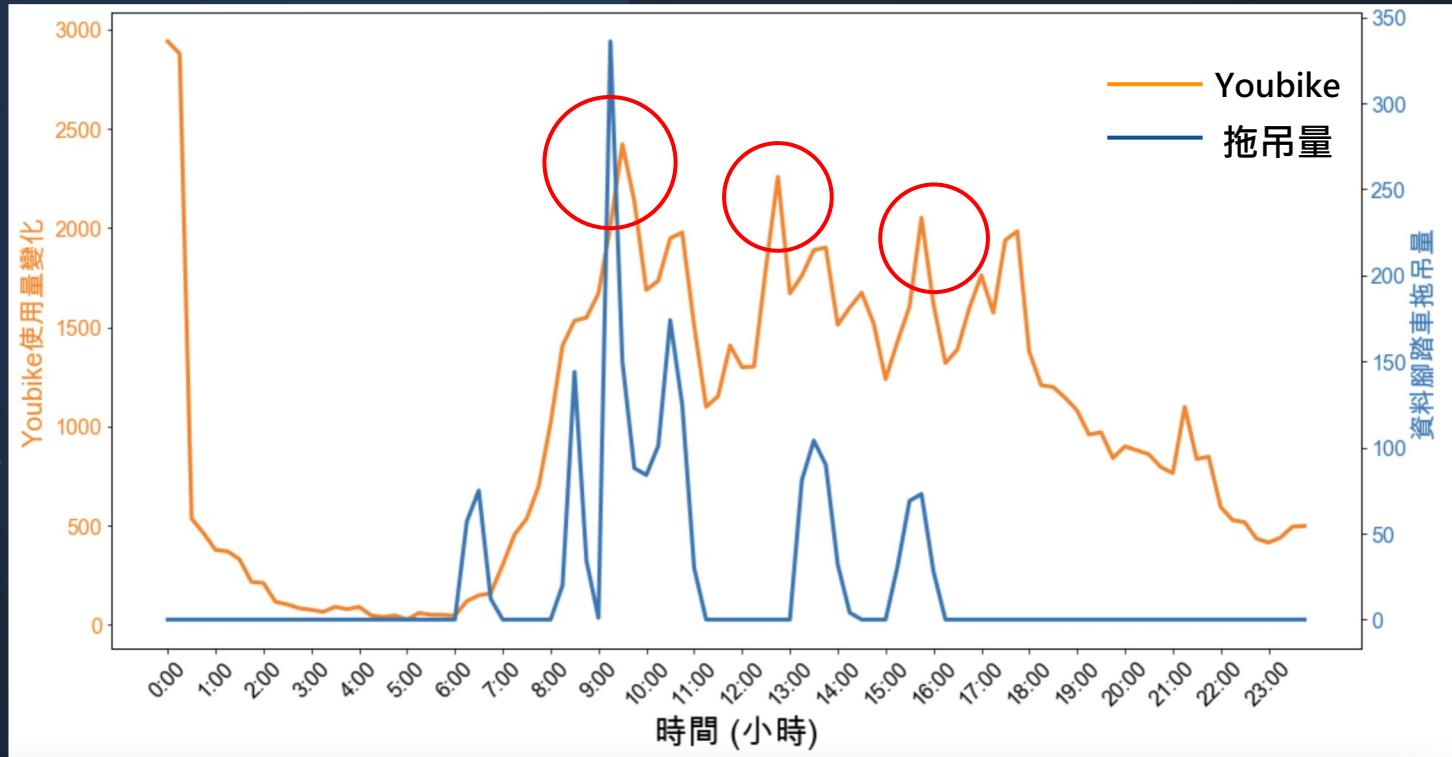
# EXPLORE THE DATA

- 歷史拖吊量與資料拖吊量比較



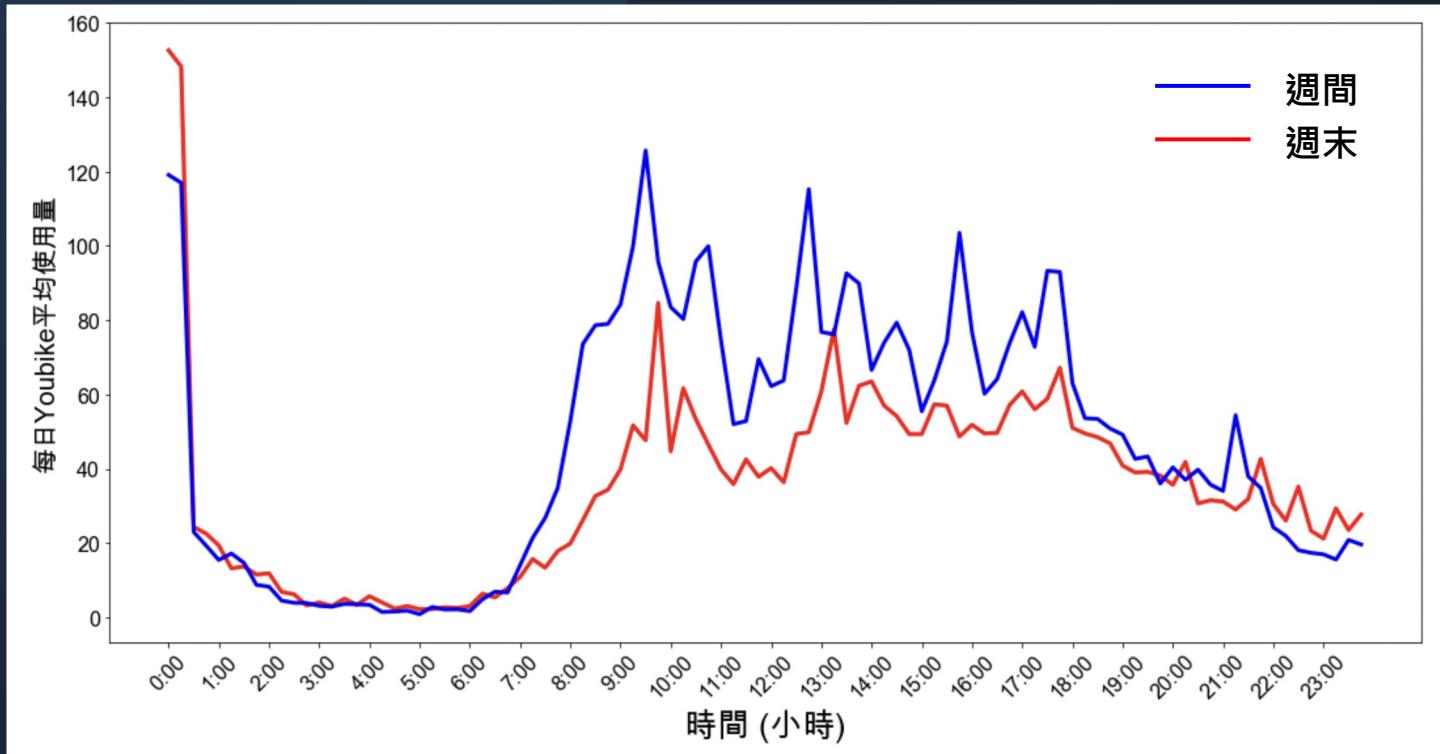
# EXPLORE THE DATA

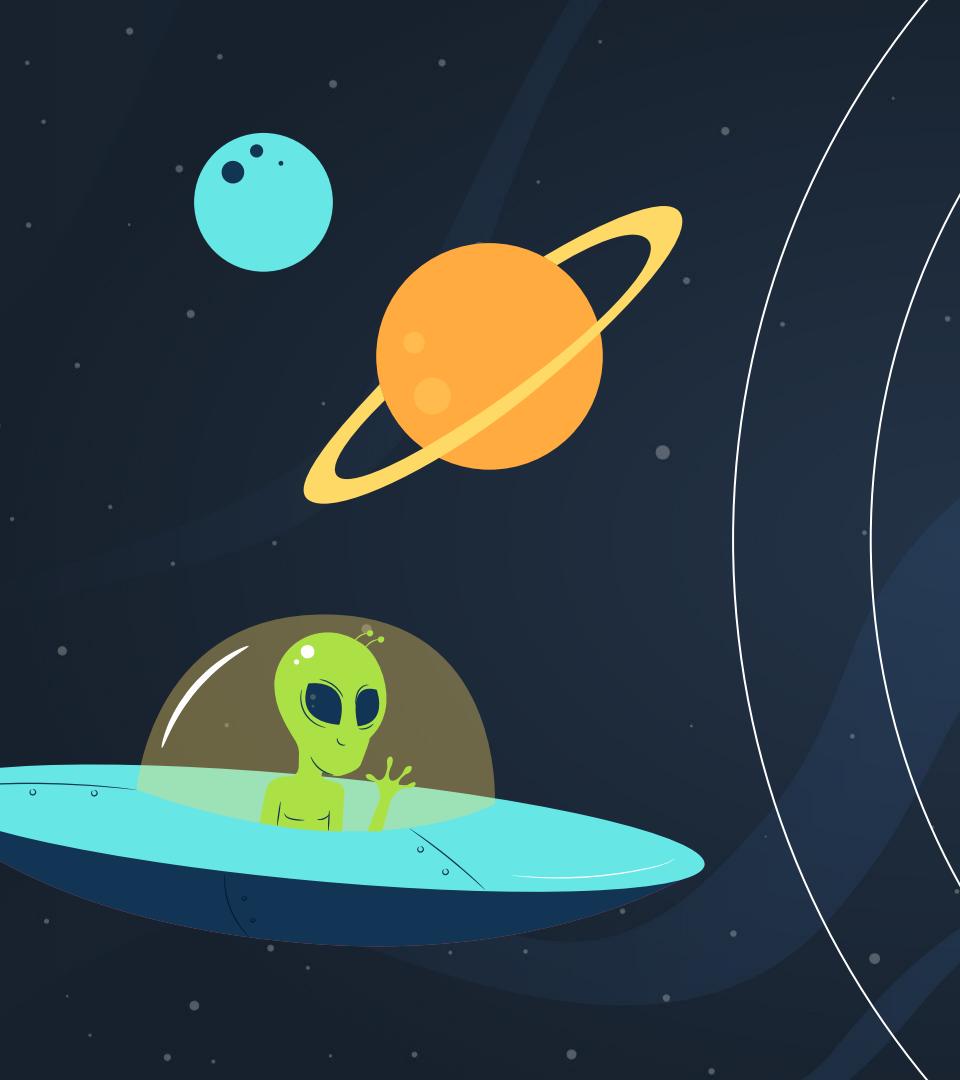
- Youbike使用量與資料拖吊量比較



# EXPLORE THE DATA

- 每日Youbike平均使用量 (週間和週末比較)





04

# DATA PREPARING

## DATA PREPARING

### □ 資料合併整理、特徵選擇(站點使用情形)

透過 YouBike 站點資料，可以取得各站點的總車位數(tot)、當下的剩餘車輛數(sbi)與剩餘空位數(bemp)

考慮到各站點在一定時間前的車輛數也可能影響使用情形，我們加入以下欄位計算各站點的車輛變化情形：

- 站點剩餘車數的延遲項(sbi\_pastN)
- 剩餘空位數的延遲項(bemp\_pastN)
- 未來N分鐘後的剩餘車輛數(sbi\_N)
- 未來N分鐘內該站點的車輛淨借出數(rent\_N)

考慮 N= 15、30、45、60 分鐘等4種情況

# DATA PREPARING

## □ 資料合併整理、特徵選擇(站點標示)

由於校內拖吊系統的地點劃分與 Youbike 站點並未完全相容，所以我們將校園內 52 個站點標註於各自距離最近的拖吊地點名稱 **place**，以及較大範圍的拖吊區域 **region**

透過計算我們可以得到該時段內各地點被拖吊的車輛數 (**pCount\_XXX**) 以及各區域被拖吊的車輛數 (**rCount\_XXX**)，其中 **XXX** 為拖吊地點或區域名稱，範圍包含拖吊資料中出現過的所有拖吊地點或區域



## DATA PREPARING

### □ 資料合併整理、特徵選擇(天氣資料)

考慮到過高或過低的氣溫、降雨等情況都可能影響使用者當下租借車輛的意願，我們選擇了該時段的氣溫(攝氏) **temp** 及降雨量(**mm**) **precip** 作為特徵欄位。

另外，在原始資料中，降雨量大於 **0** 但小於 **0.1 mm** 時紀錄為 **T**，我們於是將這些資料標示為 **0.1mm** 以進行分析。

# DATA PREPARING

## □ 資料合併整理、特徵選擇 (類別欄位處理)

將資料合併後，我們進一步將 **class** 類別的欄位做 **One Hot Encoding** 處理，這些欄位包括時間段 **time**、站點名稱 **sna**、星期幾 **weekday**、站點所屬拖吊地點 **place**、站點所屬拖吊區域 **region**

最後，我們篩選掉一些極少出現過的欄位，若一個欄位僅出現過 **104** 次或更少，代表此欄位僅出現在 **2** 個時間段 (乘以 **52** 個站點)，我們便刪去這些欄位

# DATA PREPARING

- 整理後資料範例(類別變數以One Hot Encoding 前的形式呈現 )

date	tot	sbi	countPlace	countRegion	bemp	sbi_pastN	bemp_pastN	pCount_地點	rCount_區域
2021/11/23	19	11	0.0	0.0	8	0.0	19.0	10	120
2021/11/23	10	7	0.0	0.0	3	7.0	3.0	15	140

temp	precipitation	time_時間	sna_地點	weekday	place_地點	region_區域	sbi_N	rent_N
15.5	0.1	16:00	大一女舍北側	1	研一、 大一女舍	A	10	0
15.5	0.1	12:15	社會系館南側	2	社會系館	D	15	5

# DATA PREPARING

□ 學校提供歷史拖吊資料

time	region	place	date	datetime	timerange
15:20:00	A	土木系館	2019/04/19	2019/04/19 15:20:00	1515
15:20:00	A	土木系館	2019/04/19	2019/04/19 15:20:00	1515
15:20:00	A	土木系館	2019/03/14	2019/03/14 15:20:00	1515
15:20:00	A	土木系館	2019/03/14	2019/03/14 15:20:00	1515

資料筆數：30,434 欄位：6

時間：2019年至2021年10月底

# DATA PREPARING

## 一、自變數

- 除了 `sbi_15`、`sbi_30`、`sbi_45`、`sbi_60`、`rent_15`、`rent_30`、`rent_45`、`rent_60`、`date` 之外的所有欄位

## 二、應變數

- 我們使用的 8 個應變數為 `sbi_15`、`sbi_30`、`sbi_45`、`sbi_60`、`rent_15`、`rent_30`、`rent_45`、`rent_60`，分別為該站點在 15、30、45、60 分鐘後的剩餘車數，以及該站點在未來 15、30、45、60 分鐘內的淨借出車數

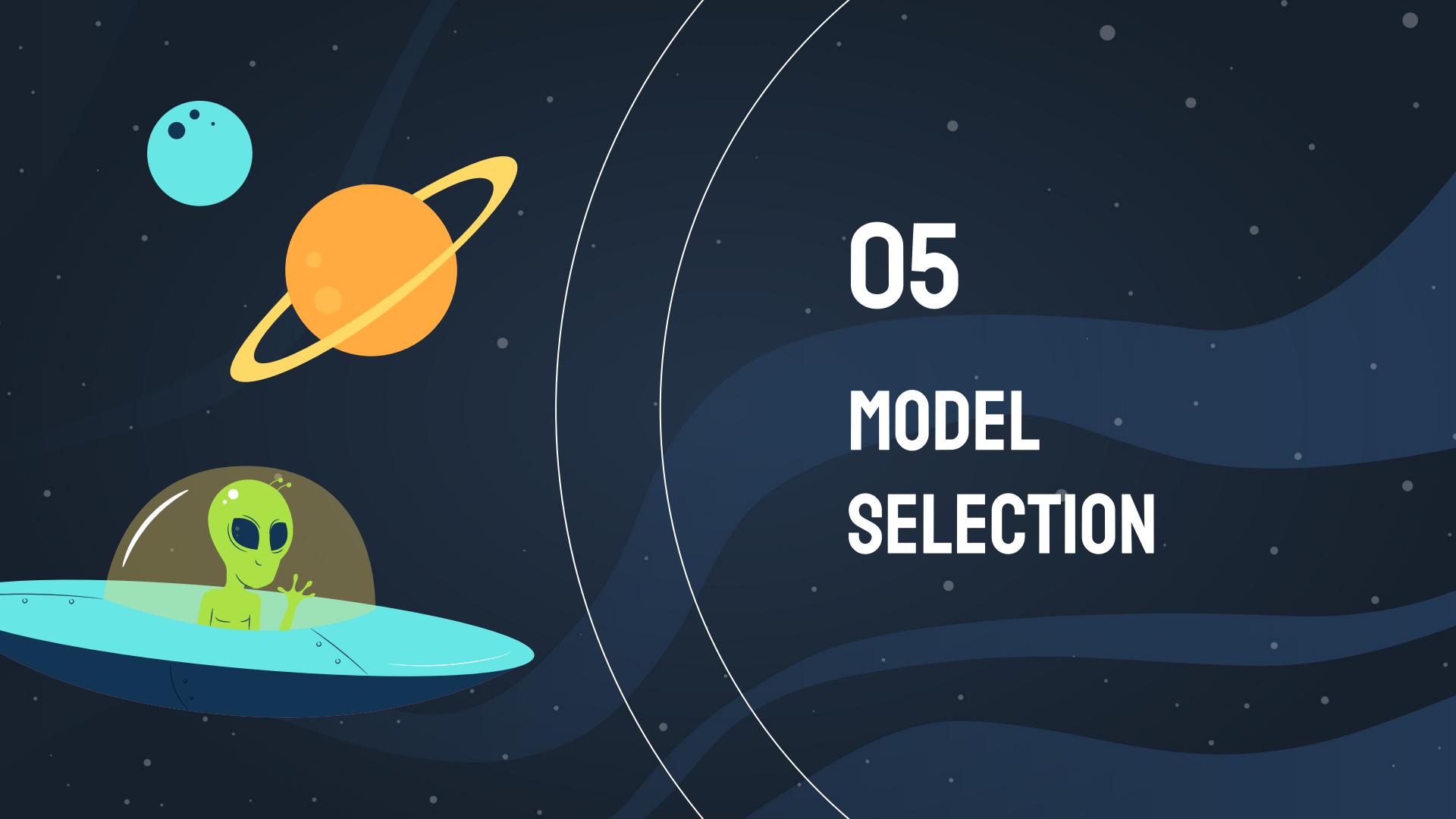
# DATA PREPARING

## 三、Train / Test 資料集

- 我們定義訓練與測試資料的比例為 7:3，並使用 `sklearn.model_selection` 中的 `train_test_split` 切分，得出擁有 74,803 筆資料的訓練集、及擁有 32,059 筆資料的測試集

## 四、模型評估

- 為了使模型預測結果在呈現上有直觀的運用價值，我們使用 **MAE**（平均絕對誤差）作為衡量指標，來呈現模型的預測效果。而在模型訓練的過程中，考量到預測誤差越大的懲罰應越大，因此使用 **RMSE** 作為訓練的誤差



# 05 MODEL SELECTION

# MODEL SELECTION

- 模型使用



OLS



Cat-boost



Decision Tree



Random Forest



MLP

# MODEL SELECTION

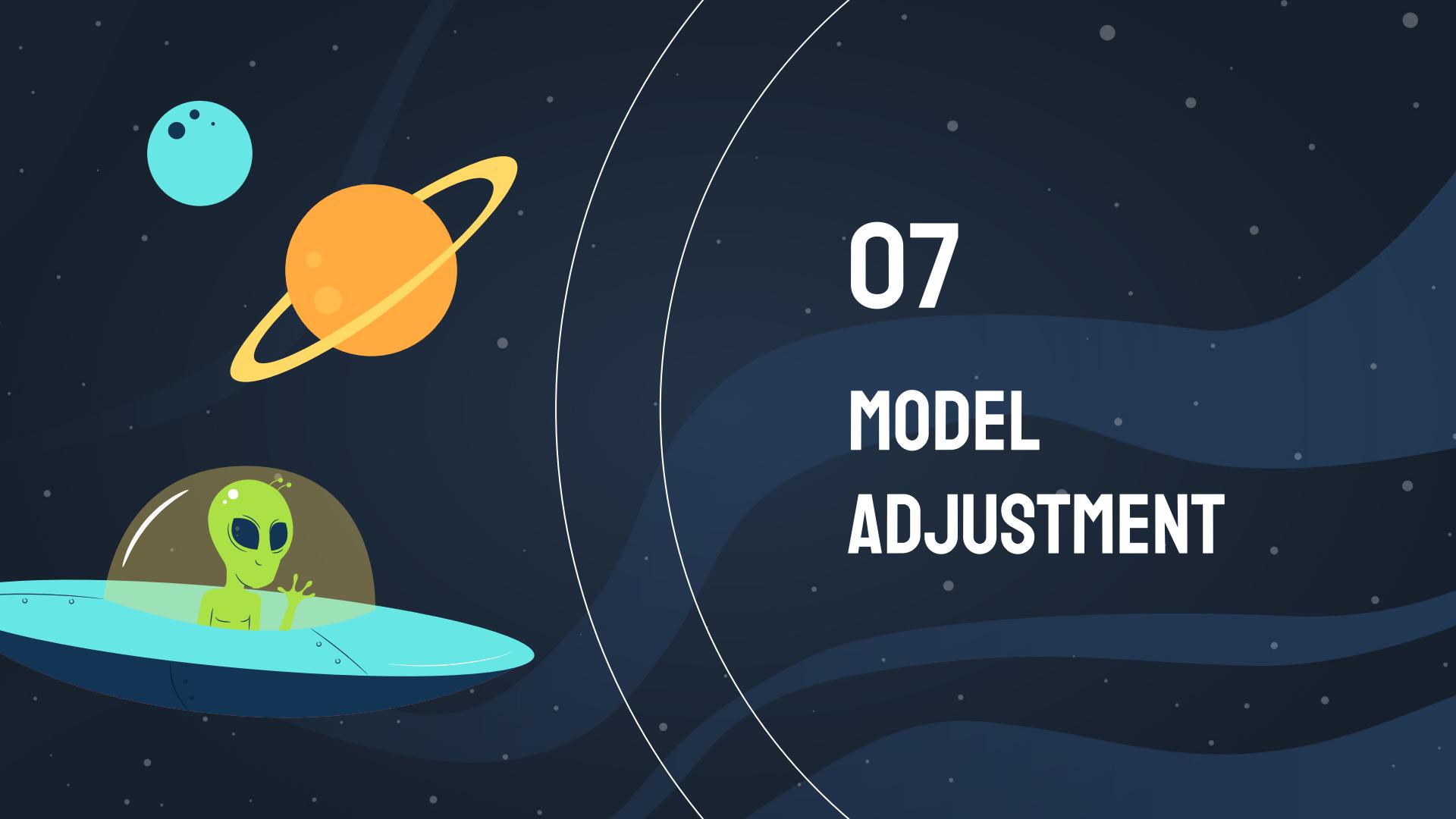
- 預測結果 所有模型的最佳 testing MAE

	sbi15	sbi30	sbi45	sbi60	rent15	rent30	rent45	rent60
OLS	1.05	1.50	1.81	2.07	1.05	1.50	1.81	2.07
catboost	1.01	1.41	1.67	1.88	1.01	1.40	1.67	1.88
Decision Tree	1.05	1.51	1.82	2.05	1.04	1.46	1.8	2.03
Random Forest	1.04	1.48	1.79	1.99	1.02	1.44	1.75	1.98
MLP	1.04	1.38	1.62	1.78	1.06	1.43	1.62	1.78

# MODEL SELECTION

## □ 預測結果 四個模型的前 10 個重要變數(除了 MLP 之外)

	OLS	CatBoost	Decision Tree	Random Forest
1	sbi	sbi	sbi	sbi
2	bemp	temp	sbi_past15	sbi_past15
3	sbi_past30	bemp	tot	sbi_past30
4	bemp_past30	sbi_past60	sbi_past30	bemp
5	sbi_past60	sbi_past15	bemp_past15	tot
6	bemp_past60	sbi_past30	bemp	sbi_past60
7	sbi_past15	sbi_past45	sbi_past45	temp
8	bemp_past15	sna_臺大大一女北側	bemp_past60	sbi_past45
9	sna_臺大檔案展示館	bemp_past60	bemp_past30	bemp_past45
10	place_土木系館	region_B	temp	bemp_past60



# 07

# MODEL ADJUSTMENT

# MODEL ADJUSTMENT

## □ MLP Model

由於Linear MLP相較於CatBoost有更好的預測效果，我們選用該模型進行Fine tuning此步驟中我們嘗試了以下做法：

由於較簡單的模型在此問題中有較好的表現，我們嘗試調整：

- 模型Hidden layer層數
- 調整模型節點數量
- 測試不同Optimizer
- 調整Mini-Batch數量
- 調整Learning rate

# MODEL ADJUSTMENT

## □ Random Forest Model

原先在訓練時是使用 **grid search**，讓模型跑過幾個參數測試哪組參數會達到最好的預測效果，然而這個方法有可能花費大量時間遍歷效果不好的參數、或是因為測試的參數不夠多而未找到最好的

為了改善這個問題，在 **fine-tuning** 的階段我們改採 **random search**

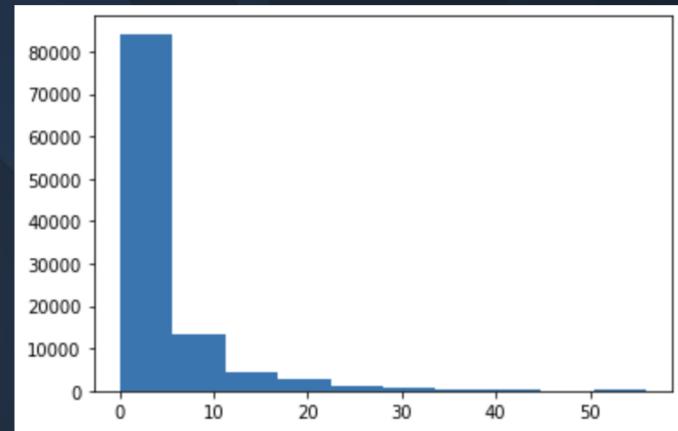
# MODEL ADJUSTMENT

## □ 資料調整

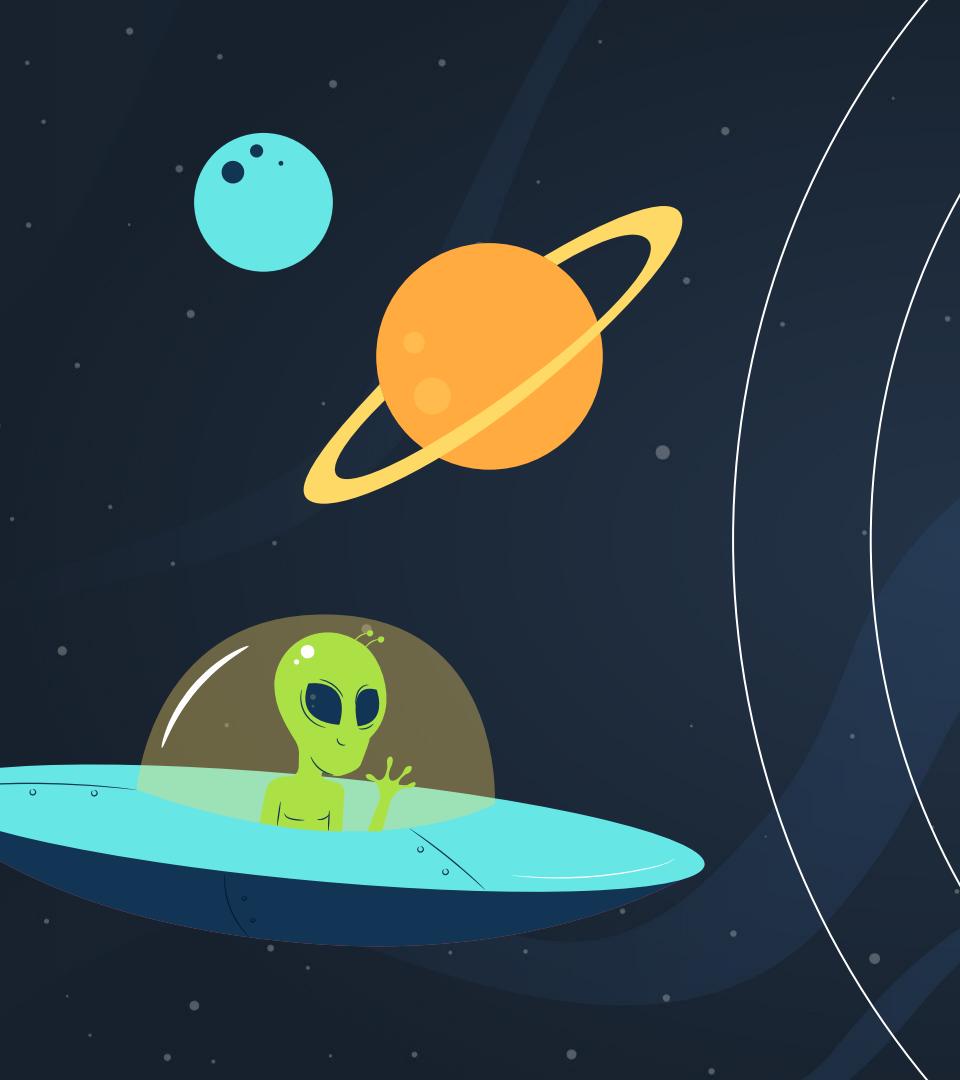
觀察預測結果及資料分布後發現，站點的可借車數之分布大部分以0到5台為主，因此模型只要全部預測0預測誤差就會很小

為避免此問題，我們嘗試利用 **SMOTE** 調整資料不平衡的狀況，讓樣本點的分布較平均

使用此方法後，訓練資料集的資料筆數增加為**1,676,370**。由右表可知使用了**SMOTE**方法後，**MAE**會較之前高，因為資料的分布變得更加均勻了，模型不再只是將大部分的資料都預測為0



	sbi15	sbi30	sbi45	sbi60	rent15	rent30	rent45	rent60
MLP	1.03	1.42	1.66	1.81	1.05	1.40	1.64	1.78
Random Forest	1.01	1.39	1.63	1.87	0.98	1.34	1.58	1.82



07

SOLUTION

# SOLUTION

- 透過模型的預測結果，我們能夠預測未來**30分鐘**時間內的**Youbike2.0**使用狀況，並使**MAE**達到**1.5**以下
- 利用預測的結果，我們能協助**Youbike**營運方有效預測未來站點的使用狀況，並藉此使得目前車輛分配的問題得到改善



# THANKS!

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.