淡江大學
大數據分析與商業智慧碩士班學程

期末報告:

Ubike 視覺化市場分析

Tamkang University
淡江大學

指導教授:高君豪

研究生:徐鵬翔

目錄

圖表目錄

壹、 研究動機與目的

一、 動機

在現今網路發達的時代，網路上充斥者許許多多看似無意義的資訊，這些看似無意義且容易被忽略的資訊經由爬蟲程式大量爬取，藉由演算法的轉換變成了有意義且可分析的資料，例如共享經濟的商業模式造就在現有運輸業者的租借模式，如 Ubike 即為此商業模式，業者作為出借方，將占點設立於街道上，並讓使用者能夠自行租借，而租借者藉由業者提供的 app 或是網站查詢可查詢各站點的可借數車輛數以及空位車輛數，而在此營運模式下所產生資訊便是可藉由大量蒐集並轉換而成為資料的資訊。藉由爬蟲程式以一固定頻率將各站點車輛數紀錄，車輛數地的變化便可可得知各站點的租借狀況，再輔以社會經濟資料為解釋變數，便可建立預測模型，用以推估運輸租借共享經濟在其他未開發市場的可行性，亦可作為車輛調度方面決策之參考。

二、 目的

本研究目的為利用紀錄 Ubike 各站點的可借車輛數推算出各站點的租借情況，並且利用此數據將租借狀況依照時間以及站點進行分類，並且利用視覺化方式來了解目前 Ubike 以設立站點的現有狀況，並在後續研究中可以依照本研究所分類的狀況進行進一步的分析亦或是建造模型進行預測，來協助調度及設立新站點的決策。

貳、 文獻探討

一、 分類方法文獻

本研究分類方式是以Hierarchical clustering(階層式集群)作為分群方式，參考Johnson, Stephen C.的"Hierarchical clustering schemes." *Psychometrika* 32.3 (1967): 241-254.。

二、 距離相似度文獻

本研究在站點之間利用Hierarchical clustering(階層式集群)分類時，參考Müller, Meinard. "Dynamic time warping." *Information retrieval for music and motion* (2007): 69-84.，利用此方法計算出各站點間的距離矩陣。

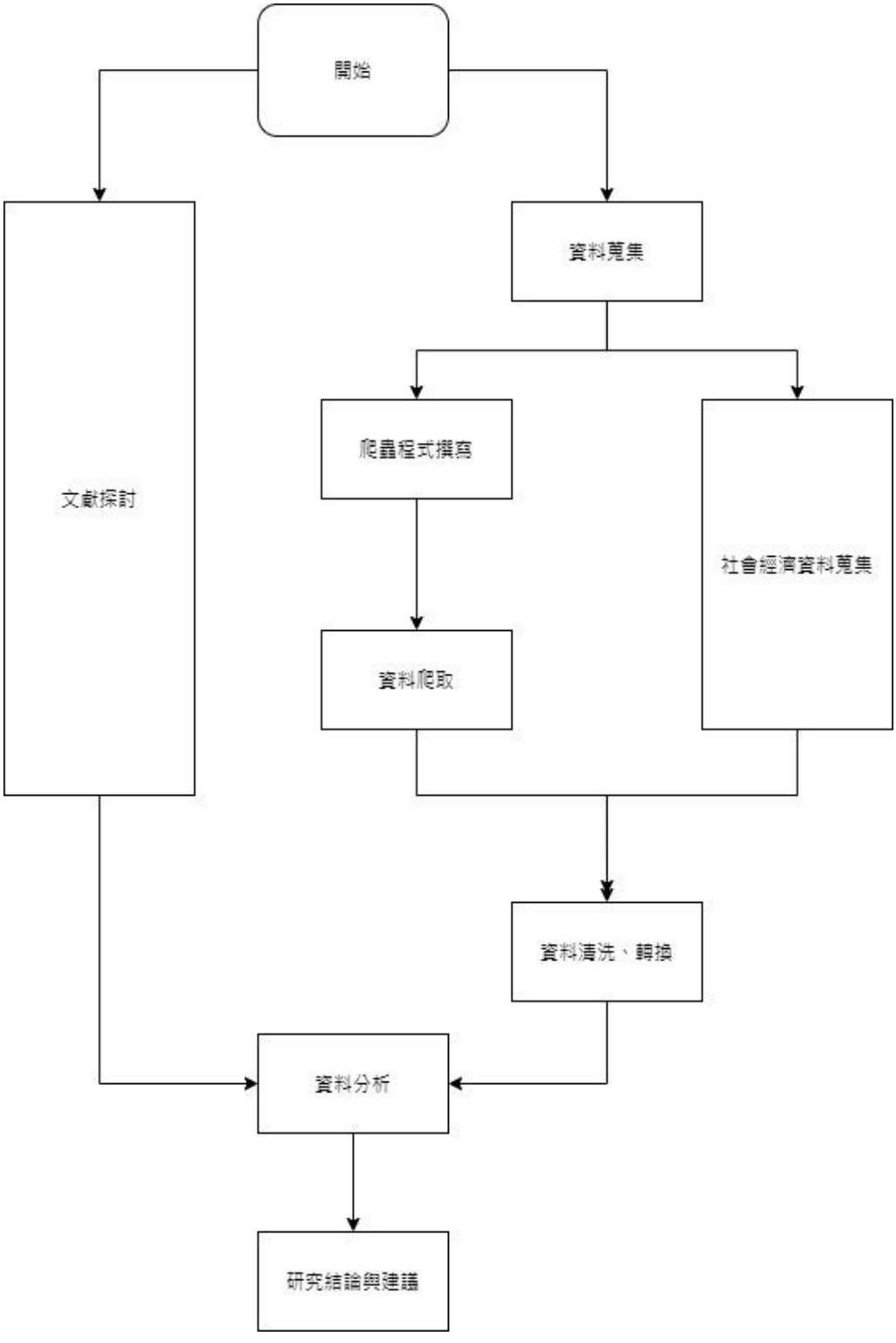
三、 視覺化呈現文獻

本研究視覺化呈現方式參考Han-Ming Wu教授以及Chun-houh Chen教授所著軟體Generalized Association Plots(GAP)，將Hierarchical clustering(階層式集群)分類後所產生樹狀結構進行排序，而本研利用此排序方法產生之序列進行上色呈現分類結果。

參、 研究流程及方法

一、 研究流程圖

圖 參.1 研究流程圖



二、 研究方法

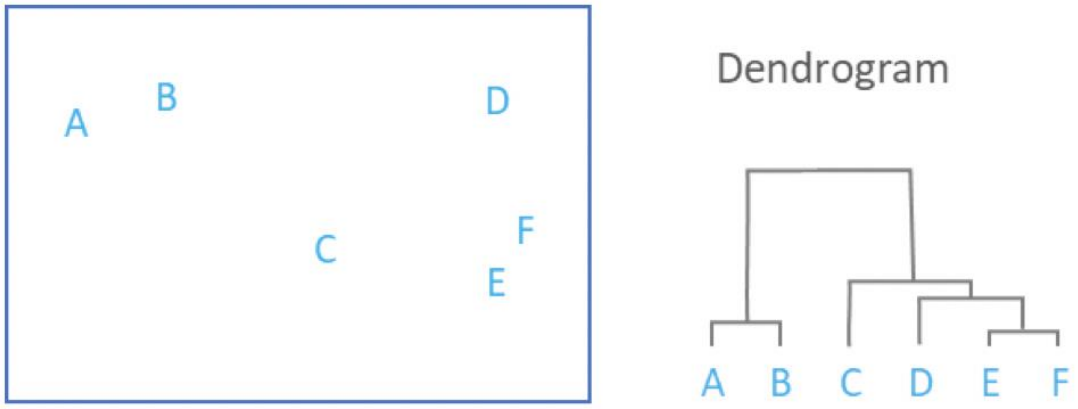
1. 資料爬取

利用爬蟲程式紀錄 Ubike 官方網站所提供 api 資料，依照每分鐘一次頻率進行紀錄，爬蟲程式碼。

2. Hierarchical clustering(階層式集群)分類

本研究所使用之 Hierarchical clustering(階層式集群)分類方法，為產生所有資料間兩兩之間的距離矩陣，並將距離矩陣內最短距離的兩筆資料合併視為一群，此群與其他資料間距離計算有、最小法、最大法以及平均法等方法，本研究所使用方法即為平均法，依照此方式不斷迭代，直至所有資料被合併為一群，將過程中進行記錄，便可得到不同群數分類的結果，其演算法數學式子為: $d(C_i,C_j)=\sum_{a\in C_i,b\in C_j} d(a,b)/|C_i||C_j|$, 其產生樹狀結果如下**錯誤！找不到參照來源。**所示，本研究之 Hierarchical clustering(階層式集群) 。

圖 參.2 Hierarchical clustering 分類示意圖



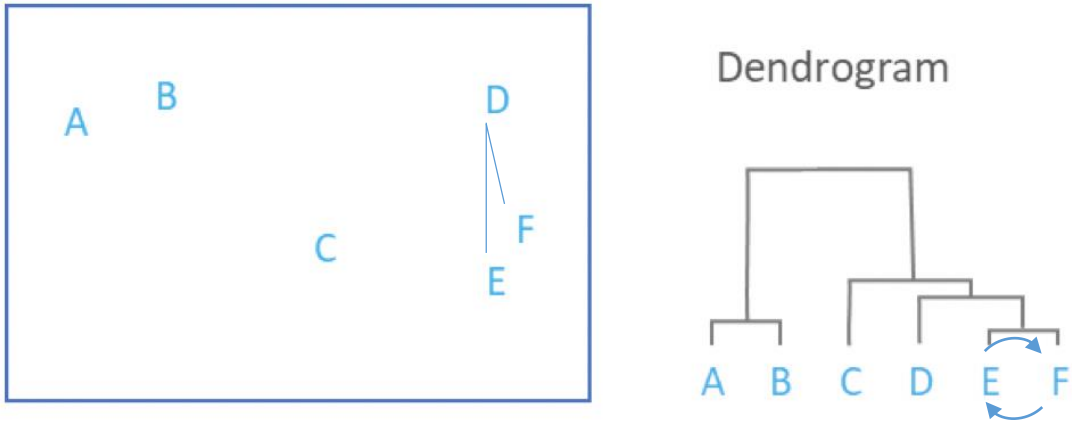
此示意圖引用至 DisplayR Blog: What is Hierarchical Clustering?, by Tim Bock

https://www.displayr.com/what-is-hierarchical-clustering/

3. Hierarchical clustering(階層式集群)排序

本研究中針對 Hierarchical clustering 分類法所產生樹狀結構結果，進一步地進行排序，針對母樹下左右兩子樹之根結點下的左右子樹(孫節點)進行排序，也就是說將左子樹的左右子樹與右子樹之距離進行比較，距離較遠的孫節點作為左子樹的左子樹，而較近的則作為左子樹的右子樹，而右子樹亦相同，依照此排序法便可產生新的樹狀結構，而此樹狀結構之葉節點便可產生一序列，在此序列中距離最近之兩節點便會排序在一起，而在此序列中排序越近意味者相似度越高，並可依照兩葉結點之距離做為上色變數並呈現，排序方式如下**錯誤! 找不到參照來源。**所示，線段 FD 相較線段 ED 較短則 E 與 F 交換位置。。

圖 參.3 Hierarchical clustering 樹狀結構排序示意圖



4. Dynamic time warping(動態時間規整)距離

本研究所使用的 Dynamic time warping 為計算兩時間序列間之距離方法，一般資料所使用的歐式距離在時間序列的資料中可能會出現波型較相近的兩資料歐式距離與波行不相似的兩資料相同的情況，因此 Dynamic time warping 方法利用動態規劃之概念計算兩時間序列資料距離，此距離能更好的表達兩波型的相似及差異，而 Dynamic time warping 方法計算數學式子如下，而本研究之 Dynamic time warping 計算程式。

$$D(i, j) = |t(i) - r(j)| + \min\{D(i+1, j), D(i+1, j+1), D(i, j+1)\}, result = D(m,n)$$

肆、 資料來源及概述

一、 資料來源

本研究所史資料來源於 Ubike 官方網站站點資料查詢所傳輸資料庫之 api:https://apis.youbike.com.tw/api/front/station/all?lang=tw&type=1，資料變數整理後如下**錯誤！找不到參照來源。**所示:

address_tw	district_tw	available_spaces	empty_spaces
中文地址	站點所在行政區	可借車輛數	可停位置術
lat	lng	name_tw	station_no
站點緯度	站點經度	站點名稱	站點編號

表 肆.1 資料變數整理

二、 資料處理

在本研究中有各站點每五分鐘記錄一次之可借車輛數，將當其可借車輛數減去前期可借車輛數，若為正數 n 則記為歸還 n 輛車輛，反之若為負數 n 則記為出借 n 輛車輛，並且依照周一至周日每小時做為單位，將歸還及出借車輛數做相加，並可得到周一至周日每小時之各站點出借以及歸還車輛數量。

三、 資料概述

1. 站點概述

本研究所蒐集資料為 Ubike 全台站點共計 2881 個站點，其位置分佈如下**錯誤！找不到參照來源。**所示，由於設備不足以計算如此龐大數量之站點距離矩陣因此在本研究中僅以台北市 503 個站點作為研究分析，其分佈如下**錯誤！找不到參照來源。**所示，由圖中可以看出目前 Ubike 集中於雙北市、桃園市、新竹市、台中市、嘉義市以及高雄市，而其他縣市目前則並未有設立站點，且站點較為集中，少數站點散布於市區外，推估為觀光景點。

台北市為我國首都，都市化程度以及人口密度較高，因此站點分怖較為平均，而未設站點區域皆為山區，應為山區較不是和腳踏車行駛因此需求較少而未設立站點，而於大安區則可以看到有站點集中在某個區域的狀況，此區域為台灣大學，應為台灣大學腹地較大且地勢平坦而大學內部禁止師生行駛動力汽機車，因此騎乘腳踏車需求較高而有此現象

圖 肆.1 全台站點分布圖

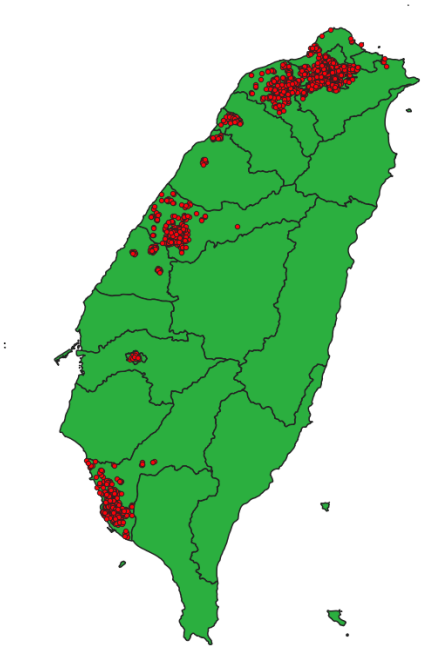
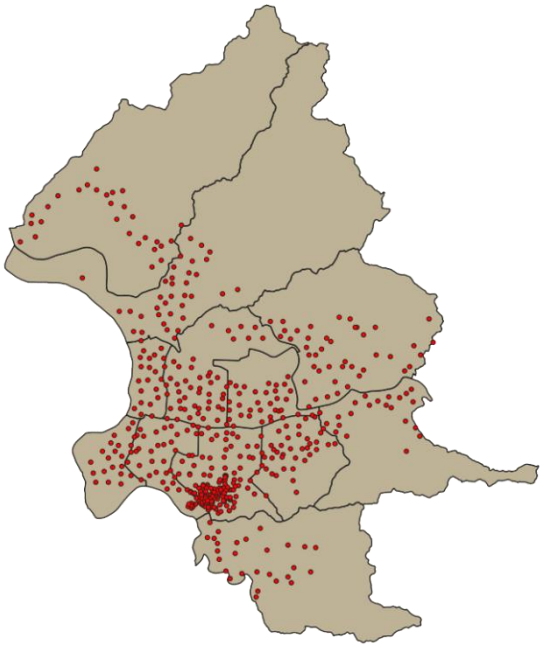


圖 肆.2 台北市站點分佈圖



2. 時間概述

本研究所蒐集資料為 12 月 23 日(星期三)21 時至 12 月 30 日(星期三)21 時共計 7 天，每五分鐘做一次紀錄，每站點資料為 2016 筆資料，經由資料處理轉換後，為周一至周日每小時一筆出借及歸還車輛數資料，每站點 168 筆資料。

伍、 研究結論

一、 依不同時間分類

本研究中計算周一至周日每小時兩兩之間以各站點出借及歸還數量作為變數利用歐式距離計算兩兩時間之距離矩陣，並以此距離矩陣執行 Hierarchical clustering 分類，產生站點樹狀結構後加以排序，可得各時間下依照相似度排序之序列，以序列中相鄰之時間歐式距離作為間距上色後以星期做為 X 軸周一至周日依序為 0 至 6，而小時做為 Y 軸由 0 點至 23 點可畫出**錯誤！找不到參照來源。**，且將星期程上 24 加上小時候做為 X 軸，各站點加總出租車輛數做為正值 Y 軸，而各站點加總歸還出租車輛數做為負值 Y 軸可繪出**錯誤！找不到參照來源。**，而依照**錯誤！找不到參照來源。**做為圖例上色，在圖例中距離越近的顏色相似度則越高。

由**錯誤！找不到參照來源。**中可以看出周一至周五的出借及歸還情況非常相似，皆是凌晨 0 點至 7 點隱約為一群於 10 點至 15 點突然轉靠近第二群，並且逐漸轉為 15 點至 19 點的第三群，20 點時突然轉為第二群並且逐漸轉向靠近第一群，重覆且規律的在各群之間轉換，而周六及周日便不太相同，在平日所出現的深紫色並未在周六出現，且 10 點至 15 點也不再靠近第二群而轉為靠近第三群，而周日部份凌晨 0 點至 10 與其他星期並未有太大不同，但周日在第三群的分布較為分散，10 點至 12 點及 16 至 19 點皆靠近第三群。

將**錯誤！找不到參照來源。**對照**錯誤！找不到參照來源。**可以看出偏向藍色系的第一群分布在出現歸還總量較低的狀況，因此以總量來看第一群的出租及歸還數量屬於較低的一群，而第二群較為適中，而第三群為歸還數量高峰，但在紫色屬於藍色系中最極端值，若以此結論來看總量應為最低，但在圖中卻可發現其總量僅次於紅色系的第三群，推估為其雖總量與紅色系相似但各站點出借量及歸還量可能與紅色系相反，且其在時間上分布於 8 點至 10 點為上班上課時間，紅色系的第三群則分布於 15 點至 18 點的下班下課時段，也非常合理。

藉由本章節之分析可以初步了解出借及歸還量在時間上的變化趨勢，但各站點間的變化狀況並須更進一步的由各站點進行分析。

圖 伍.1 時間分類周曆圖

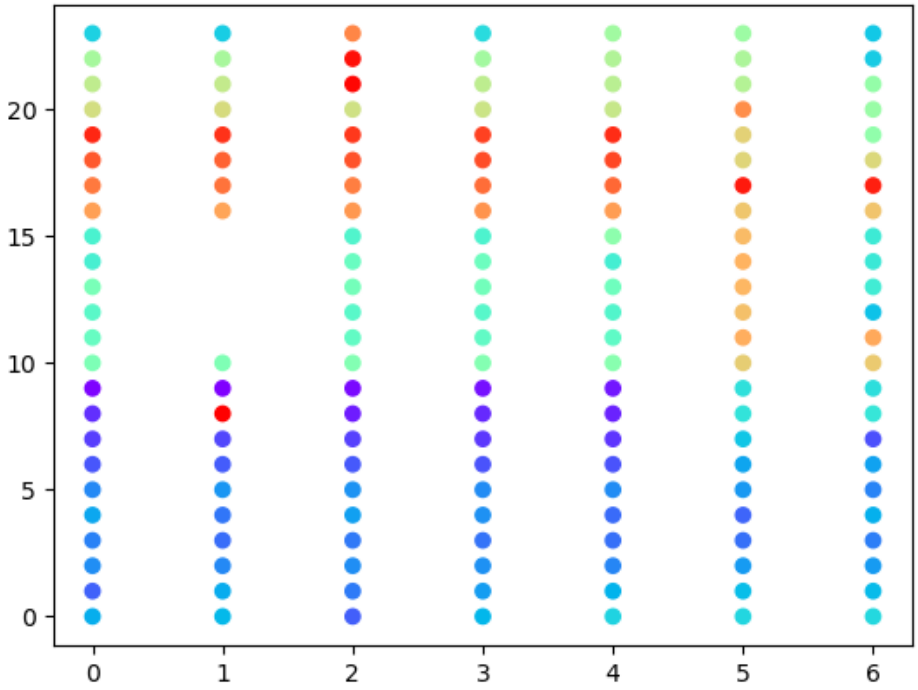


圖 伍.2 出租歸還總量折線圖

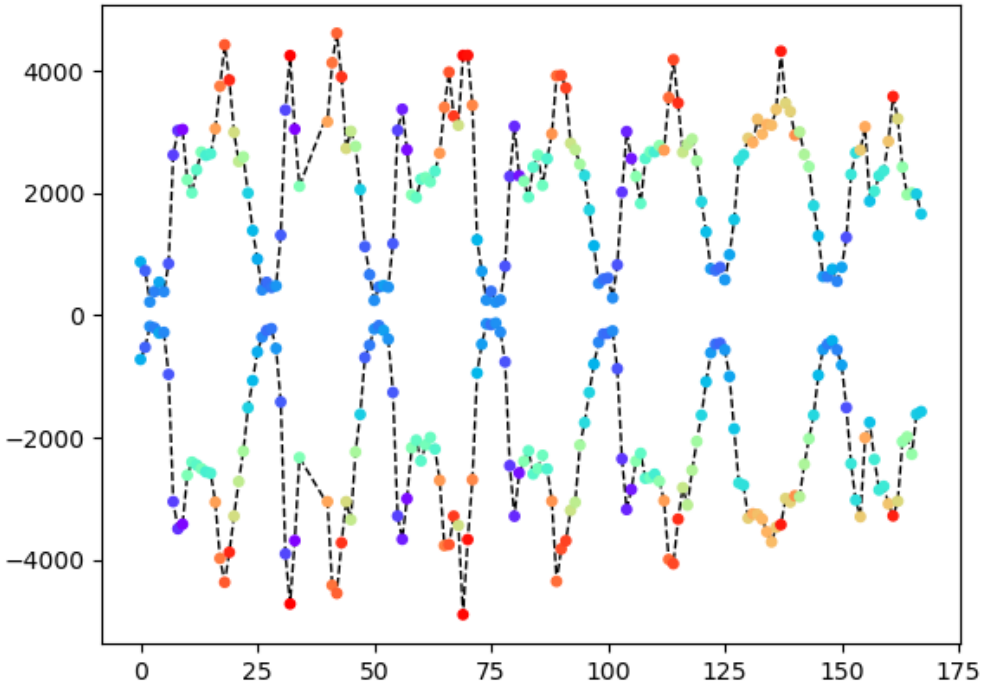


圖 伍.3 序列上色圖例

二、 依不同站點分類

本研究中依各站點不同時間做為變數以 Dynamic time warping 方法計算出各站點兩兩站點間之距離矩陣，並且以 Hierarchical clustering 方法分類，與前述時間分類相同方法進行排序得到各站點之序列，並也依照相鄰兩站點距離做為間距進行上色，將結果利用 GIS 繪至於地圖上如**錯誤！找不到參照來源。**，並且以**錯誤！找不到參照來源。**做為對照，除此之外也將各站點之出借及歸還量繪出折線圖中如**錯誤！找不到參照來源。**。

由**錯誤！找不到參照來源。**中可以發現各類站點除在台灣大學之外並未有明顯的群聚情況，且分布平均，因此可推估大部份各站點之出借及歸還情況未受特殊地點而影響，而是如生態圈一般散佈，但台灣大學內部的各站點間出借歸還情況即明顯相似度較高，對照**錯誤！找不到參照來源。**可發現各站點之尖離峰時段並未有明顯不同，但藍色系站點類則出借歸還數量在尖離峰有較大的差異，而紅色系類站點則較為平緩且穩定，而介於兩者間的色系在出借及歸還情況也介於兩者之間。

圖 伍.4 台北市行政區對照圖

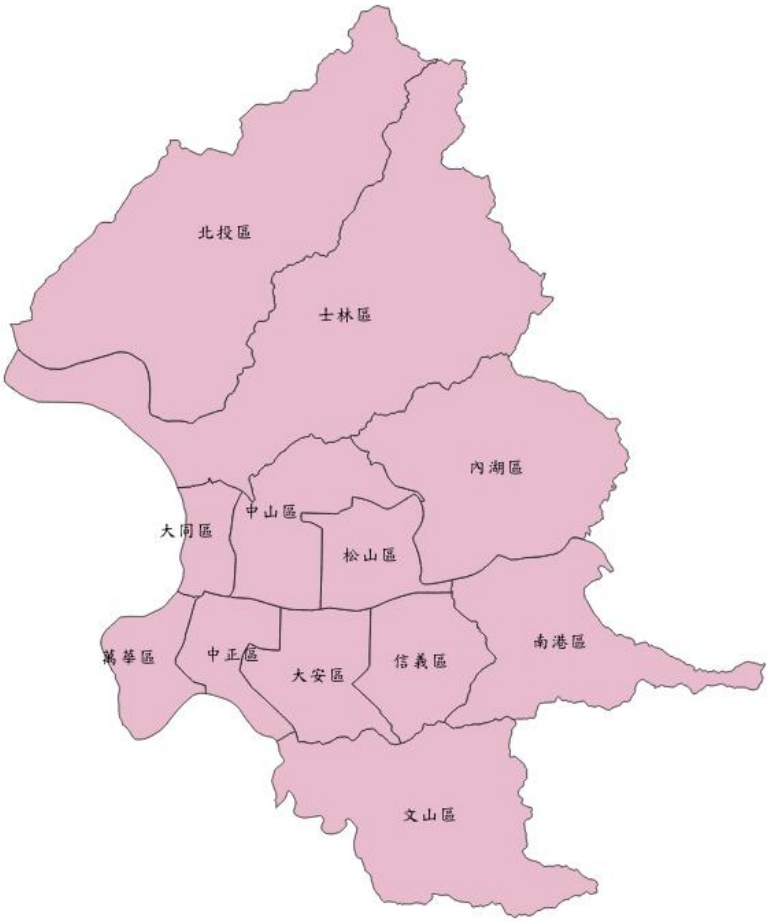


圖 伍.5 各站點分類分佈圖

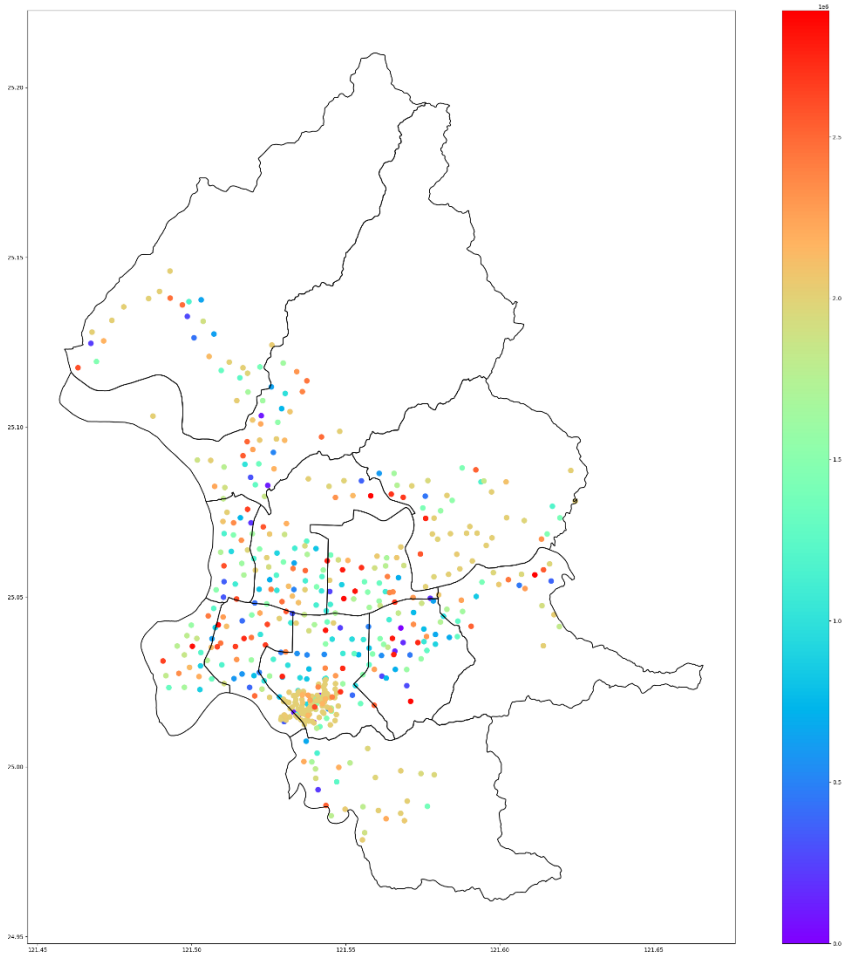
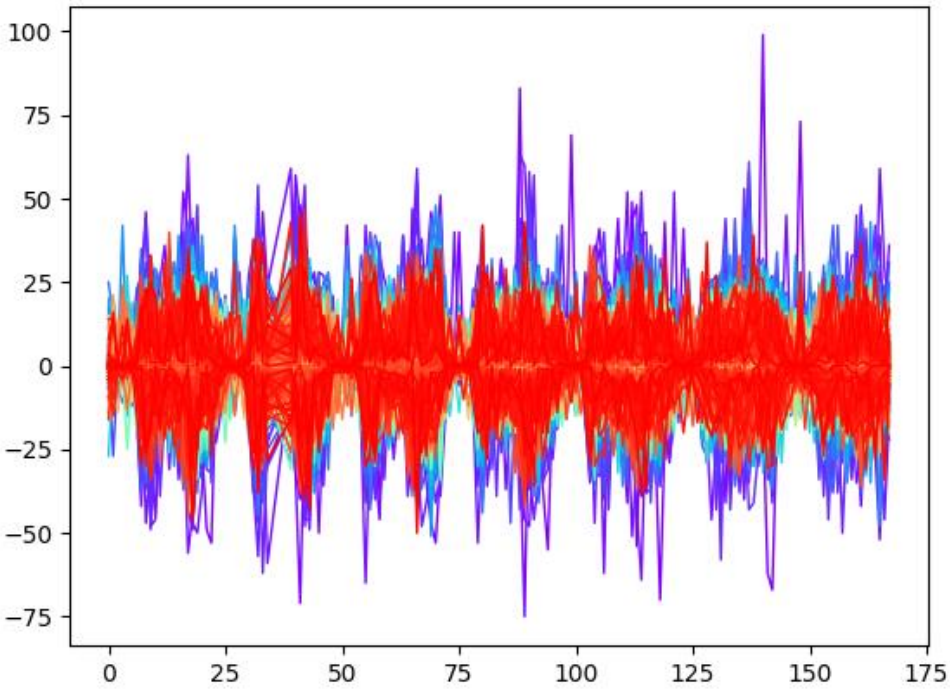


圖 伍.6 各站點出租及歸還量折線圖



陸、 結論及後續研究建議

本研究於最後一章節將結論進行整理如下結論，且整理此次研究不足可以改進的部分，及可接續之研究方向整理如下

一、 結論

1.由時間分類結果可以看出出借以及歸還情況擁有規律的周期性。

2.平日(周一至周五)與假日在出借及歸還情況上有不同情形

3.由站點分類結果可以得知相近的站點出借歸還情況並不會較相似，且不同類型站點如生態圈一般均勻分佈在地圖上，但某些特殊區域會造成相近站點出借歸還型態相似，如台灣大學內站點。

4.各站點尖離峰時段相似。

5.所有研究資料、程式碼及相關成果皆以上傳至 GitHub 供檢視參考:
https://github.com/P-Xiang-flying/Ubikev-market-analyize.git

二、 後續研究建議

1.本研究因時間及設備原因並未使用所有站點進行分類，若使用全部站點進行分類，或許可以看出不同縣市是否會有不同的出借歸還情況

2.本研究所使用時間資料恰好為一周，若能蒐集長的時間段，便能將不同日期的周一至周日每小時出借及歸還量做平均，取得更接近常態的資料，結果可能會更加地接近常態。

3.本研究僅將站點及時間分別進行分群，若能依照站點分群後，針對不同群組進行時間分群，應能更好的了解各站點的出借與歸還情形，反之先依照時間分群，並進行細部的站點分群亦然。

4.本研究在分群時在描述出借及歸還時皆以量的形式進行計算，並未考量到站點本身的大小，若考量站點本身大小，或許在分群結果會更偏向波型的不同而不是量的多寡。

5.本研究最終結果僅以圖片呈現，若能製作成互動式介面或系統，應能更好的進行分析及判斷。

6.本研究中除死用站點出借及歸還資訊外並未使用其他資料輔助，若蒐集其他社會經濟資料亦或是其他點位資料，如:7-11 點位資料、捷運站點位資料等……也許能利用決策樹或是迴歸分析等方法更好的了解各站點分群結果原因，也能做在日後再進行新站點設置位置以及究站點退場時做為參考。

7.經由本研究之站點及時間分群後，可針對各群進行 Queueing Theory 分析，以及庫存管理分析，藉此了解各群站點現況，能更好的了解各站點車位是否不足或過多，以及在車輛調度上能做到更好的規劃。