2019

中華郵政大數據競賽

全民來家郵

郵政業務量預測與地區分群

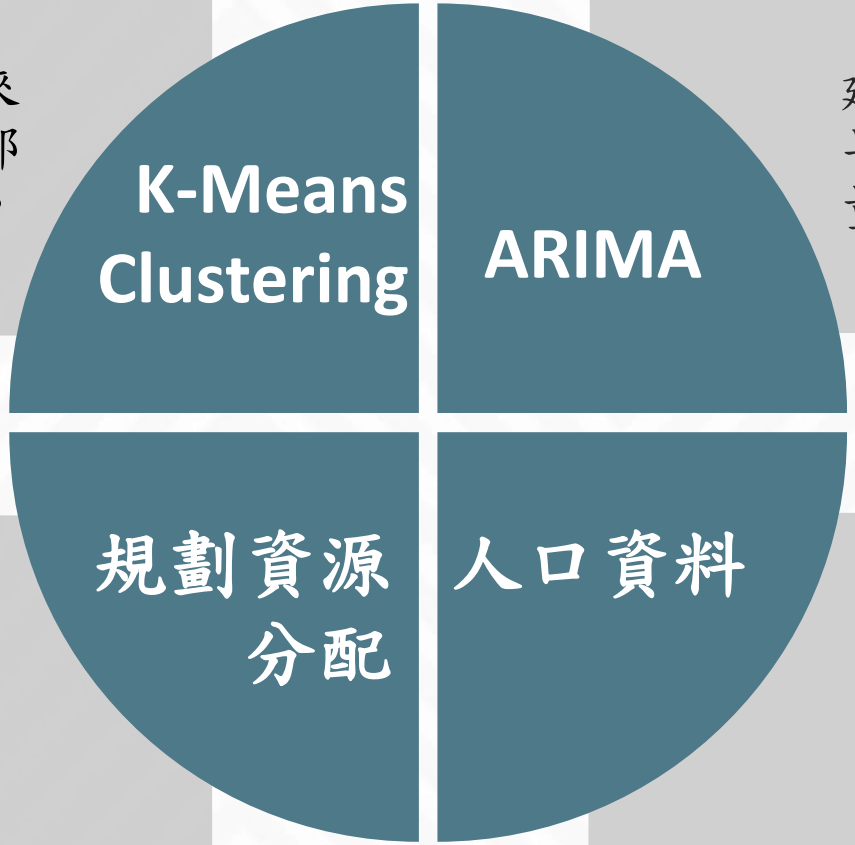
下一隊

指導老師：謝邦昌 老師

目錄
CONTENT

摘要

做出集群分類，發現未來郵局做決策時，應以六都和非六都作為決策區分。



建立業務量預測，可以提早安排人力，使人均處理量相同。

利用現有數據規劃資源分配，例如i郵箱據點等等。

透過全國人口資料，篩選出人口數較多村里，將郵務車路徑特別經過此區域的全家，也可以增設i郵箱來分擔郵局處理量。

提案動機

ONE

提高郵務處理效率

透過提早預測業務量，進行郵務人員調動，以支配即將配送包裹量。

TWO

郵局與全家合作

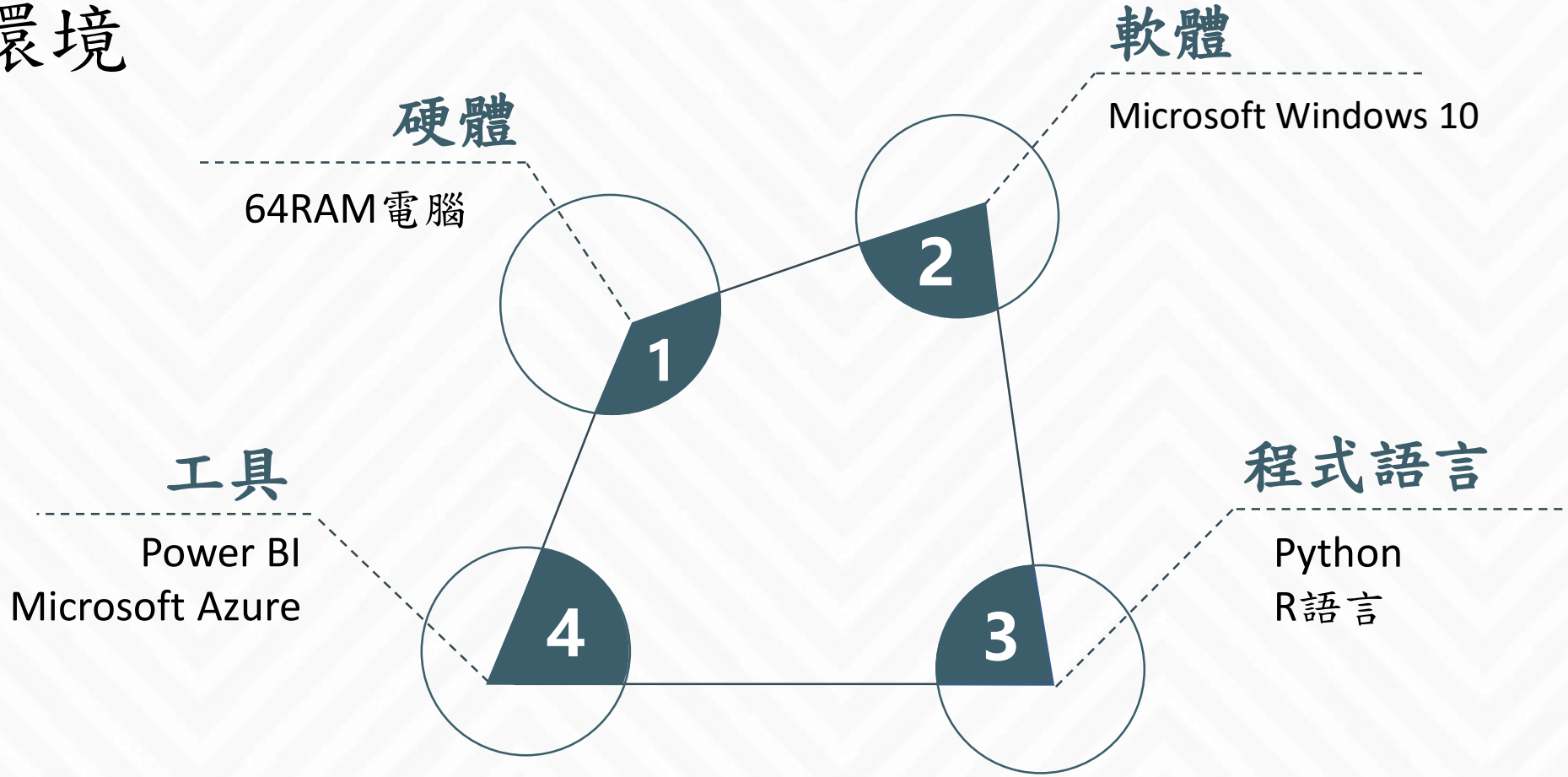
郵局與全家的強強合作，使我們產生一個發想，以全家分佈位置，提供郵務車路徑參考，讓全家補足郵局不足的區域。

THREE

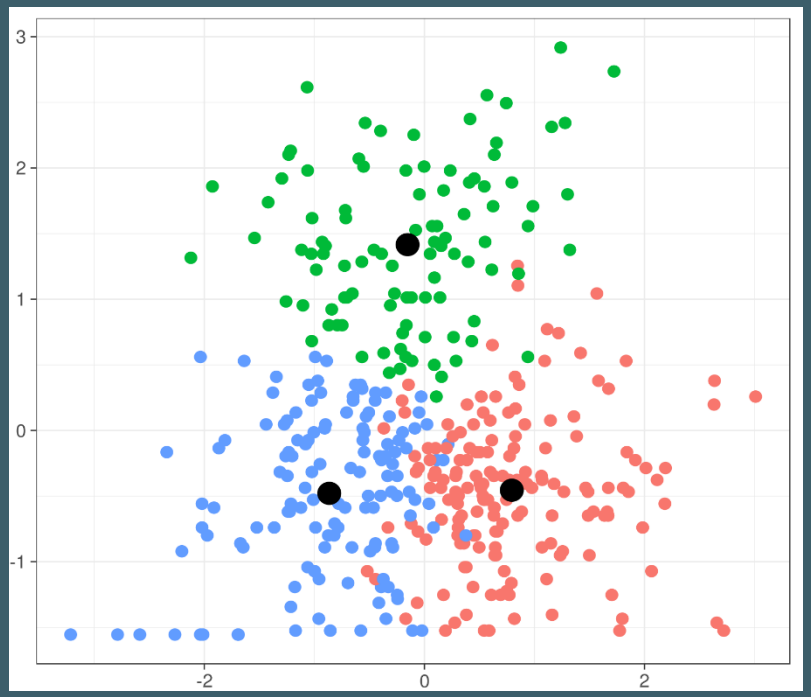
幫助未來決策方向

統整各地人口分佈，郵務車數、郵局數、商業發展等等，並將台灣分群，以利決策。

處理環境


硬體
64RAM電腦
軟體
Microsoft Windows 10
1
2
工具
Power BI
Microsoft Azure
程式語言
Python
R語言
4
3

K-Means Clustering



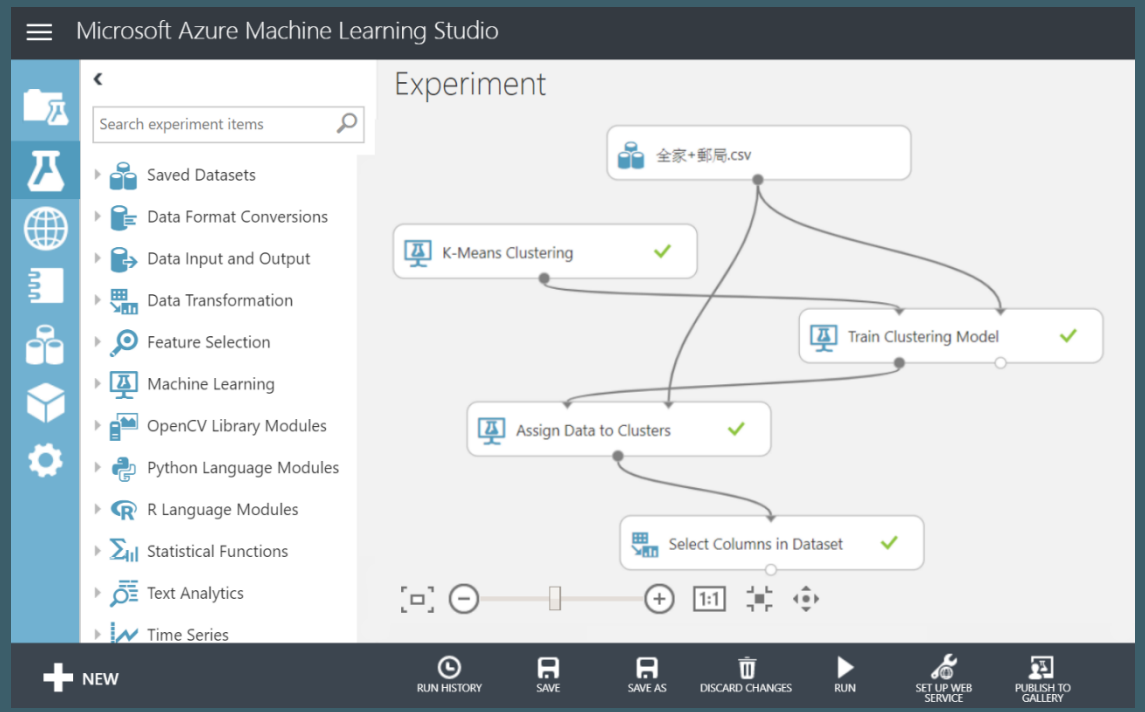
• K-Means概念圖

重複直到所有群心不在有太大的變動(收斂)

放入變數	設定分群	計算距離	判別群組	更新群組
將變數投射在平面上做分群	初始設定k個分類，並隨機選擇該分群的群心	計算每筆資料與各群心之歐式距離	將每筆資料分類到最近的群心之群組	當加進一筆資料到群組裡，便更新一次新的群心

K-Means 集群分析的概念就是「物以類聚」，是非監督式學習(Unsupervised learning)，意為得到的資料沒有任何Ground truth，必須透過資料本身去調整分群。

K-Means Clustering



• Microsoft Azure Machine Learning 介面

K-Means工具

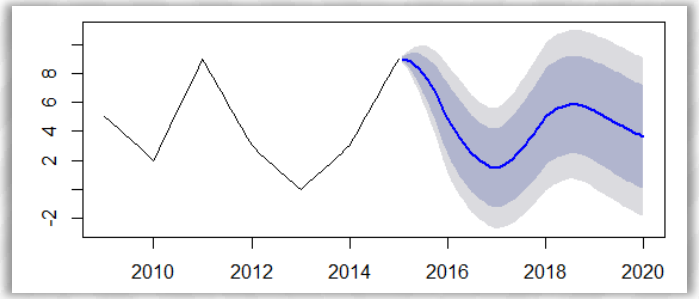
Microsoft Azure Machine Learning Studio

Microsoft Azure Machine Learning Studio採用容易使用的瀏覽器為基底，完全無須撰寫任何程式碼，按幾下就能將構想化成部署。可用來在雲端中快速建置、部署及共用預測性分析解決方案。

我們只要將資料匯入Microsoft Azure Machine Learning ，提交指令給機器執行，便能計算出最佳分群結果。

ARIMA

ARIMA (p,d,q) 稱為差分自回歸滑動平均模型，
為時間序列預測分析方法之一



AR為「自回歸」，p為自回歸階數
MA為「滑動平均」，q為滑動平均階數
d為時間序列成為平穩時所做的差分次數

ARIMA (p,d,q) 模型先對階數非平穩的資料歷史數據Yt進行d次差分處理，得到新的平穩的數據序列Xt，將Xt擬合ARMA (p,q) 模型，然後再將原d次差分還原，便可以得到Yt的預測數據。

分析流程

設定目標

預測業務量以及地區分群。

資料採集

郵政內部數據結合外部資料，包含全家、人口、工廠、商店、公司分布，並運用地理資訊圖資雲服務平台轉換經緯度。

資料清洗

運用R語言對資料預處理，以便建模。

數據建模分析

透過ARIMA時間序列和K-Means建模分析。

分析結果

郵政業務量預測與地區分群。

地區分群－資料採集與清洗

全家地址

使用R語言將全家官網上的地址爬取下來，並將其地址轉為經緯度

郵局地址

政府開放資料平台

人口分布

中華資料採礦協加上全國達康利用行政院主計處普查資料推估

郵務車分布

GPS檔結合郵局地址，得知郵務車所在縣市

投遞失敗率

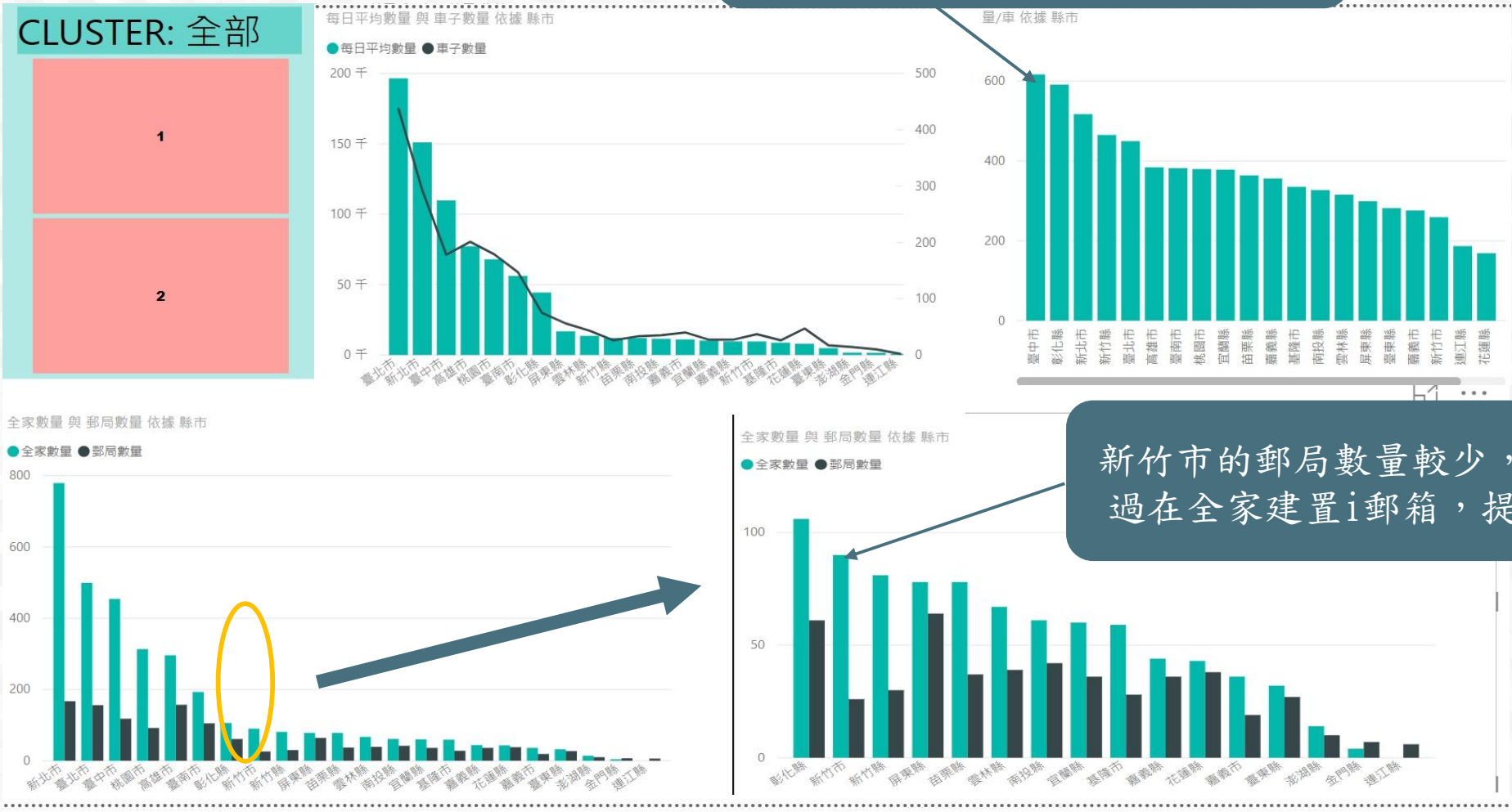
TTS檔中擷取投遞成功和失敗資料，並結合郵局地址將其分出縣市，加以運算

工廠、商店、公司

中華資料採礦協會節結合全國達康推估

地區分群 – 資料發現

臺中市和彰化縣應增加郵務車以免造成每車平均郵務量過高



新竹市的郵局數量較少，但是可透過在全家建置i郵箱，提高普及率

地區分群 – 資料發現

全國人口數前五大里	人口數
高雄市左營區福山里	42927
高雄市左營區菜公里	34645
高雄市左營區新上里	30362
高雄市鼓山區龍水里	25208
高雄市楠梓區清豐里	25150

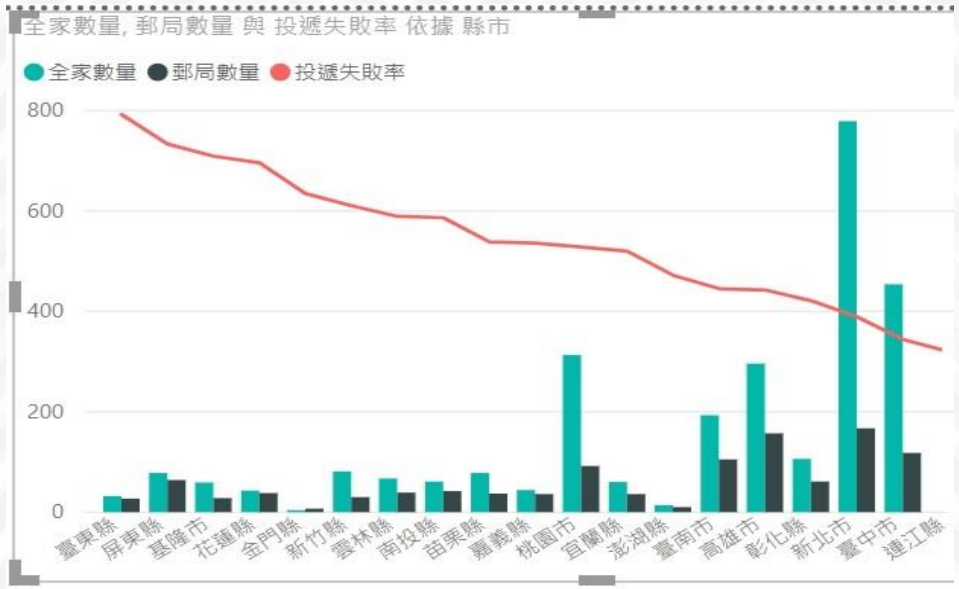
福山里的全家位置

高雄市左營區福山里58鄰重愛路208號
高雄市左營區福山里9鄰華夏路1761號
高雄市左營區福山里60鄰華夏路1168號
高雄市左營區福山里38鄰榮佑路67號
高雄市左營區福山里51鄰榮總路225號

圖中為福山里的全家和郵局分佈，可發現這麼大的里，只有2家郵局，可藉由全家來補足;郵務車路徑規劃及i郵箱也可設置在此。

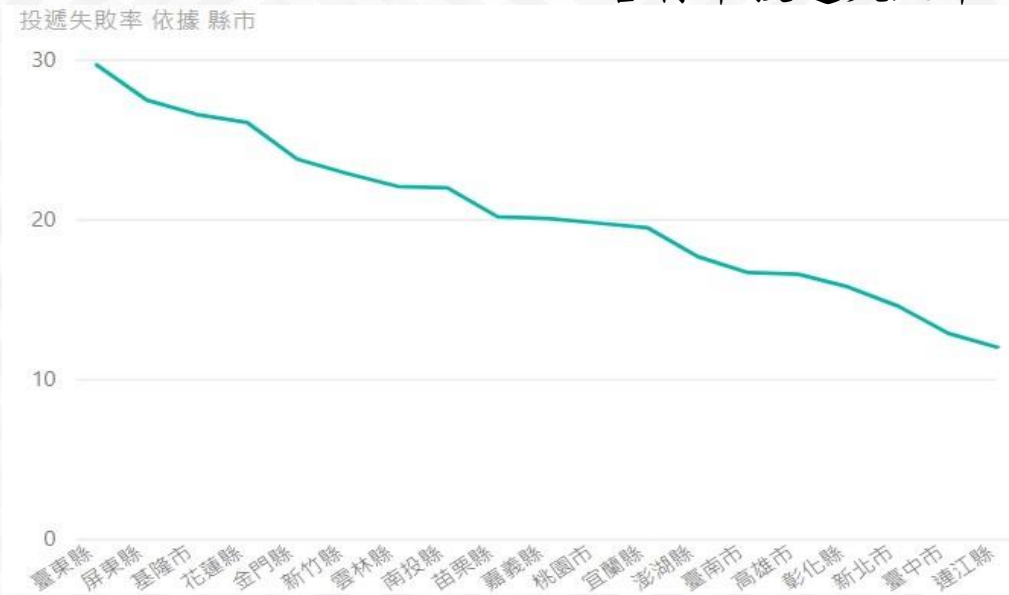
地區分群 – 資料發現

全家、郵局數量跟投遞失敗率



投遞失敗率較高的地方可以藉由全家和郵局共同設置i郵箱，居民領取時間不受限制

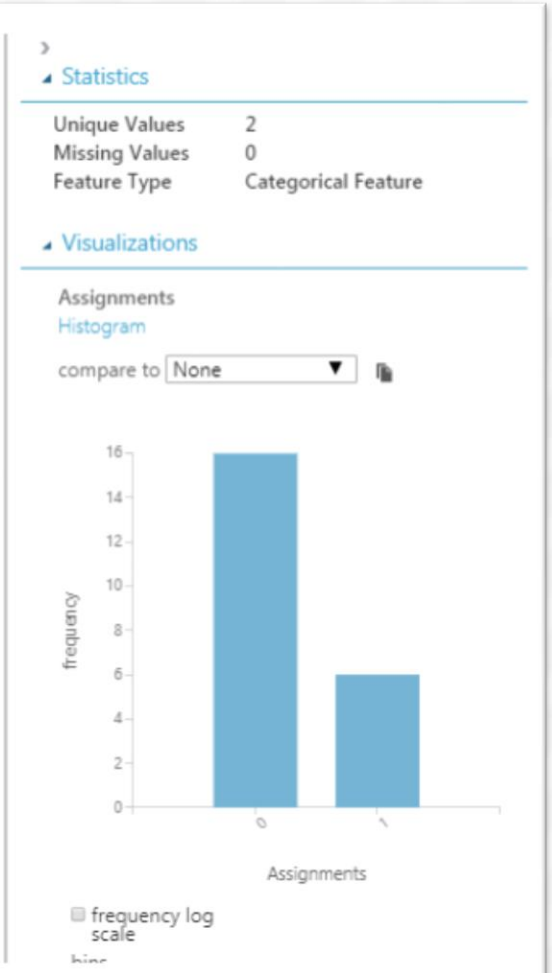
各縣市投遞失敗率



投遞失敗率高的地方主要在非六都，可能原因為從事行業的類型及居住型態

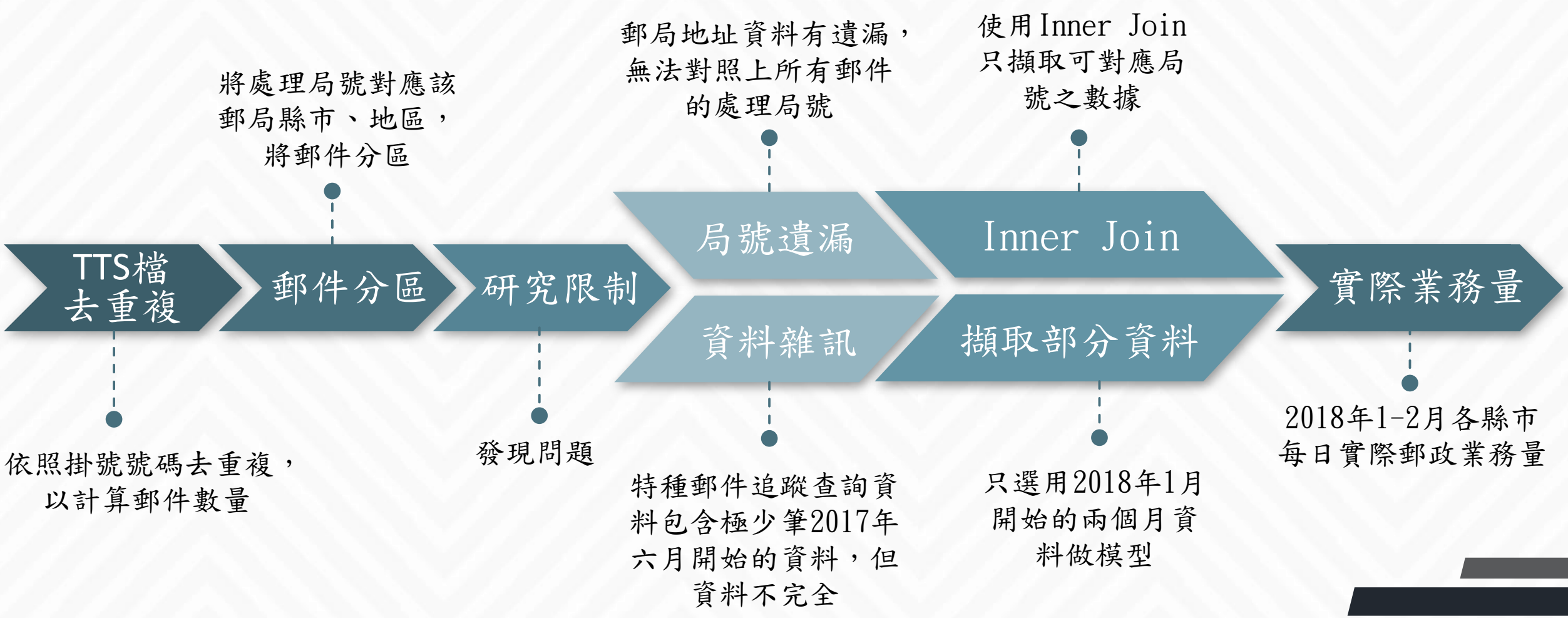
地區分群 – 分析結果

縣市	全家數量	郵局數量	人口數量	車子數量	投遞失敗率	工廠數	商店家	公司數	Assignments
南投縣	61	42	507210	35	22	684	17980	5437	0
苗栗縣	78	37	561495	33	20.2	1742	16257	5116	0
雲林縣	67	39	697105	43	22.1	1841	19420	6351	0
彰化縣	106	61	1288803	75	15.8	9507	36714	24684	0
臺中市	454	118	2754191	178	12.9	17072	81219	85892	1
宜蘭縣	60	36	457808	27	19.5	977	22084	5958	0
桃園市	313	92	2124339	179	19.8	10924	64730	54220	1
基隆市	59	28	371837	26	26.6	267	13541	6025	0
新北市	779	167	3972204	292	14.6	20295	120574	132081	1
新竹市	90	26	435238	37	0.4	1316	17363	10912	0
新竹縣	81	30	544624	26	22.9	1721	14782	9811	0
澎湖縣	14	10	102329	14	17.7	103	3161	435	0
金門縣	4	7	133185	10	23.8	89	10517	723	0
臺南市	193	105	1885388	147	16.7	9039	65510	33624	1
嘉義縣	44	36	517339	27	20.1	1642	11698	3801	0
嘉義市	36	19	270254	40	0	435	11310	4379	0
高雄市	296	157	2778092	201	16.6	7330	106808	77757	1
屏東縣	78	64	838584	56	27.5	1209	22610	6329	0
臺東縣	32	27	221504	17	29.7	200	8752	1387	0
花蓮縣	43	38	331372	47	26.1	389	15243	3714	0
臺北市	499	156	2702925	437	0	1203	86006	185925	1
連江縣	0	6	12536	2	12	2	320	94	0



將全家數量、郵局數量、人口數量、車子數量、投遞失敗率、工廠數、商店數、公司數匯入 Microsoft Azure Machine Learning，計算出的分群(如左圖)，恰好分出六都與非六都。

業務量預測 - 資料清洗

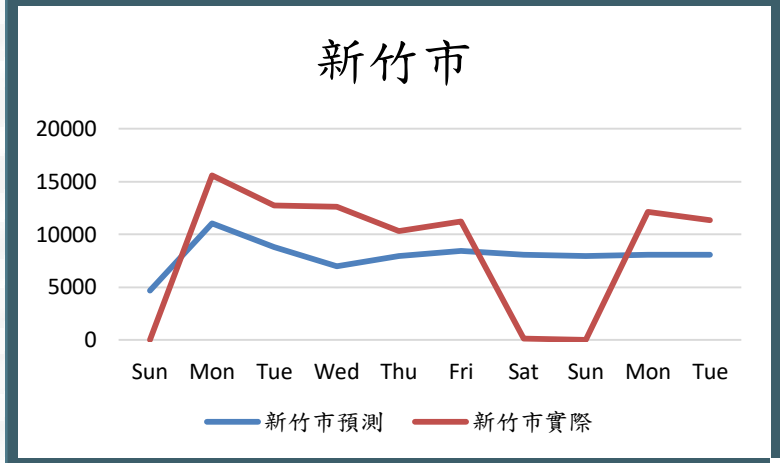

TTS檔
去重複
依照掛號號碼去重複，
以計算郵件數量
郵件分區
將處理局號對應該
郵局縣市、地區，
將郵件分區
研究限制
發現問題
局號遺漏
郵局地址資料有遺漏，
無法對照上所有郵件
的處理局號
資料雜訊
特種郵件追蹤查詢資
料包含極少筆2017年
六月開始的資料，但
資料不完全
Inner Join
使用Inner Join
只擷取可對應局
號之數據
擷取部分資料
只選用2018年1月
開始的兩個月資
料做模型
實際業務量
2018年1-2月各縣市
每日實際郵政業務量

業務量預測－調整模型

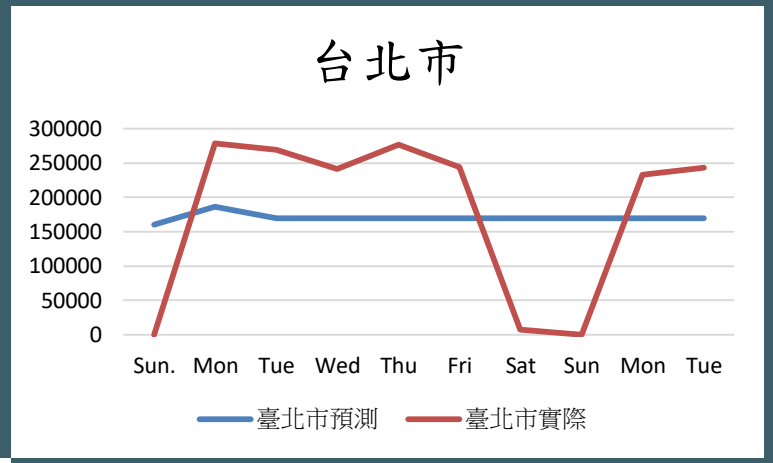
均方誤	高雄市	花蓮縣	基隆市	嘉義市	嘉義縣	金門縣	連江縣	苗栗縣	南投縣	澎湖縣	屏東縣
59天	60729.9	5477.504	6340.904	7745.272	7028.792	1071.549	275.3558	8470.449	8057.513	1062.837	11813.9
63天	44770.01	4070.233	4880.158	6833.327	5171.652	1044.682	232.0883	6517.305	6400.556	1047.199	8977.246
均方誤	臺北市	臺東縣	臺南市	臺中市	桃園市	新北市	新竹市	新竹縣	宜蘭縣	雲林縣	彰化縣
59天	142357.6	3136.975	43505.94	86661.42	51150.54	111141.5	6969.583	8352.504	7095.231	9672.656	35658.24
63天	114377.6	2934.852	30410.74	59928.82	35620.64	83895.38	5055.367	6332.854	5775.613	7261.724	23290.1

原先只選用2018年1月開始的2個月總共59天的資料做模型，但發現業務量以一星期為一個週期，所以我們將模型略做調整，改為以63天的資料做模型，其均方誤便降低許多。另外，我們也試過許多方式調整模型，例如將實際值取log或標準化再預測，最後還是以原始值預測的均方誤為最小。

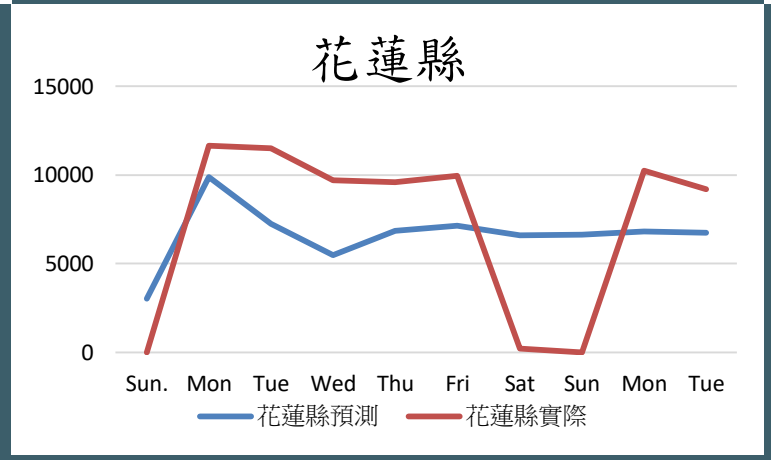
業務量預測 - 預測結果



以63天的資料做為模型，
預測未來十天的業務量



僅有一季資料，其中又遇二月年假，使非常態性資料佔據不小比例，導致預測偏差



資料不足，故預測值很快就趨向固定值

總結

地區分群
01

透過K-Means Clustering將台灣分成六都與非六都，未來可運用分群結果決策。

業務量預測
02

透過ARIMA時間序列預測業務量，以方便進行郵政人員調動，提高效率。

分析建議
03

臺中市和彰化縣可增加郵務車，新竹市可在全家建置i郵箱。

未來期望
04

補足遺漏值，分析可以縮小到鄉鎮。資料量從季節變分年，提高準確性。且期望未來能做到預測送達時間。

資料來源

行政院主計總處統計：https://www.dgbas.gov.tw/np.asp?ctNode=2824
中華資料採礦協會：https://www.cdms.org.tw/
全國達康：http://www.trend-go.com/
地理資訊圖資雲服務平台：https://www.tgos.tw/TGOS/Web/TGOS_Home.aspx
全家官網：https://www.family.com.tw/Marketing/index.aspx
模型理論：
https://docs.microsoft.com/zh-tw/azure/machine-learning/service/concept-azure-machine-learning-architecture
https://medium.com/@chih.sheng.huang821/機器學習-集群分析-k-means-clustering-e608a7fe1b43
https://wiki.mbalib.com/zh-tw/ARIMA模型

The End