資料探勘期中報告

R10525069 林子傑

主題：clustering

問題

- 根據缺席紀錄，將缺席時間依照程度分類
  - 員工 ID
  - 缺席原因
  - 缺席時間
  - 通勤距離與費用
  - 工作表現
  - 身體狀態
  - ……

資料選擇

- 缺席原因
- 通勤狀況
- 健康狀態
- 工作表現

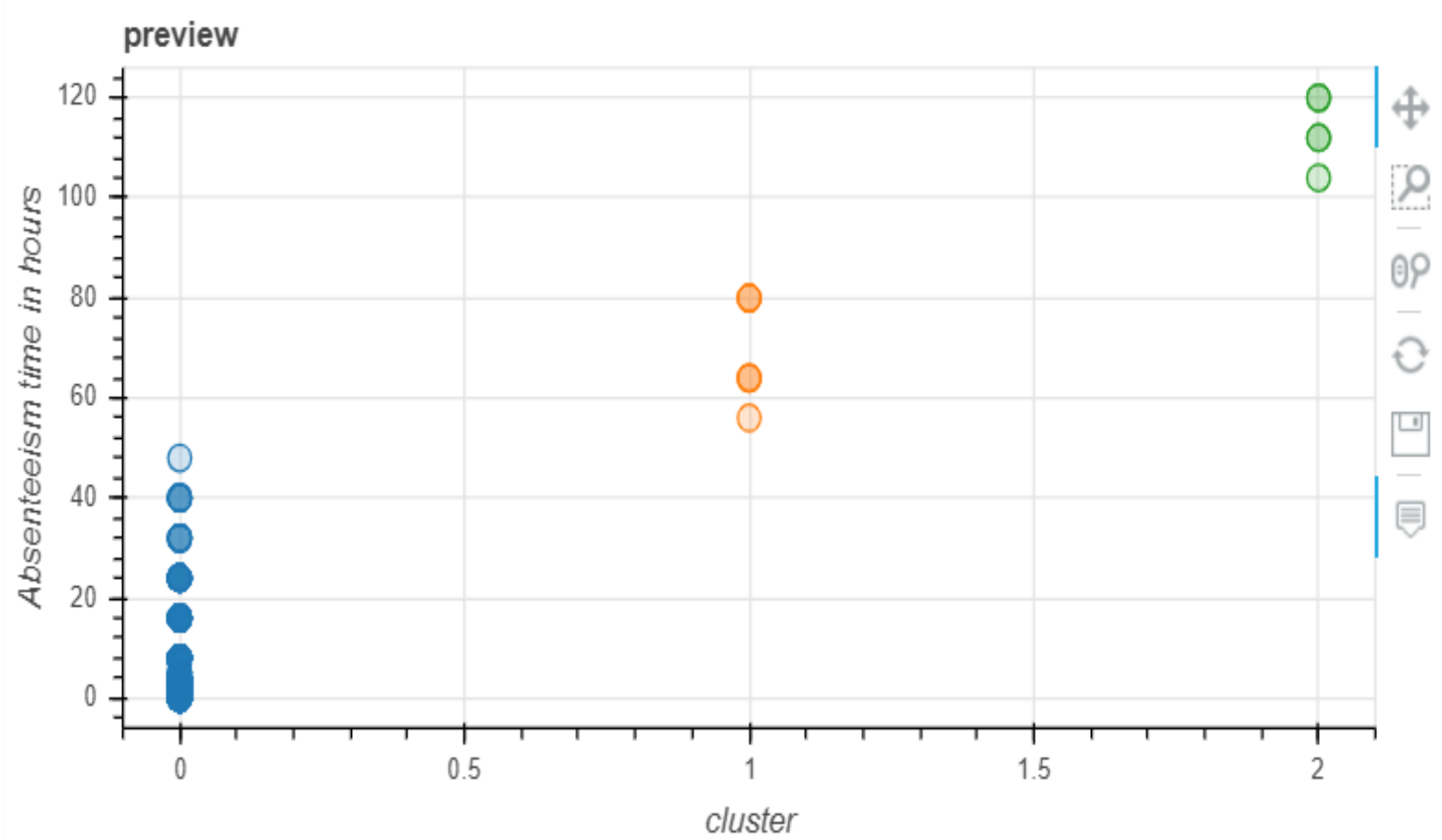
資料前處理

- 對沒有固定選項且分布範圍廣的資料做 Normalize 和 Flitering Missing Value

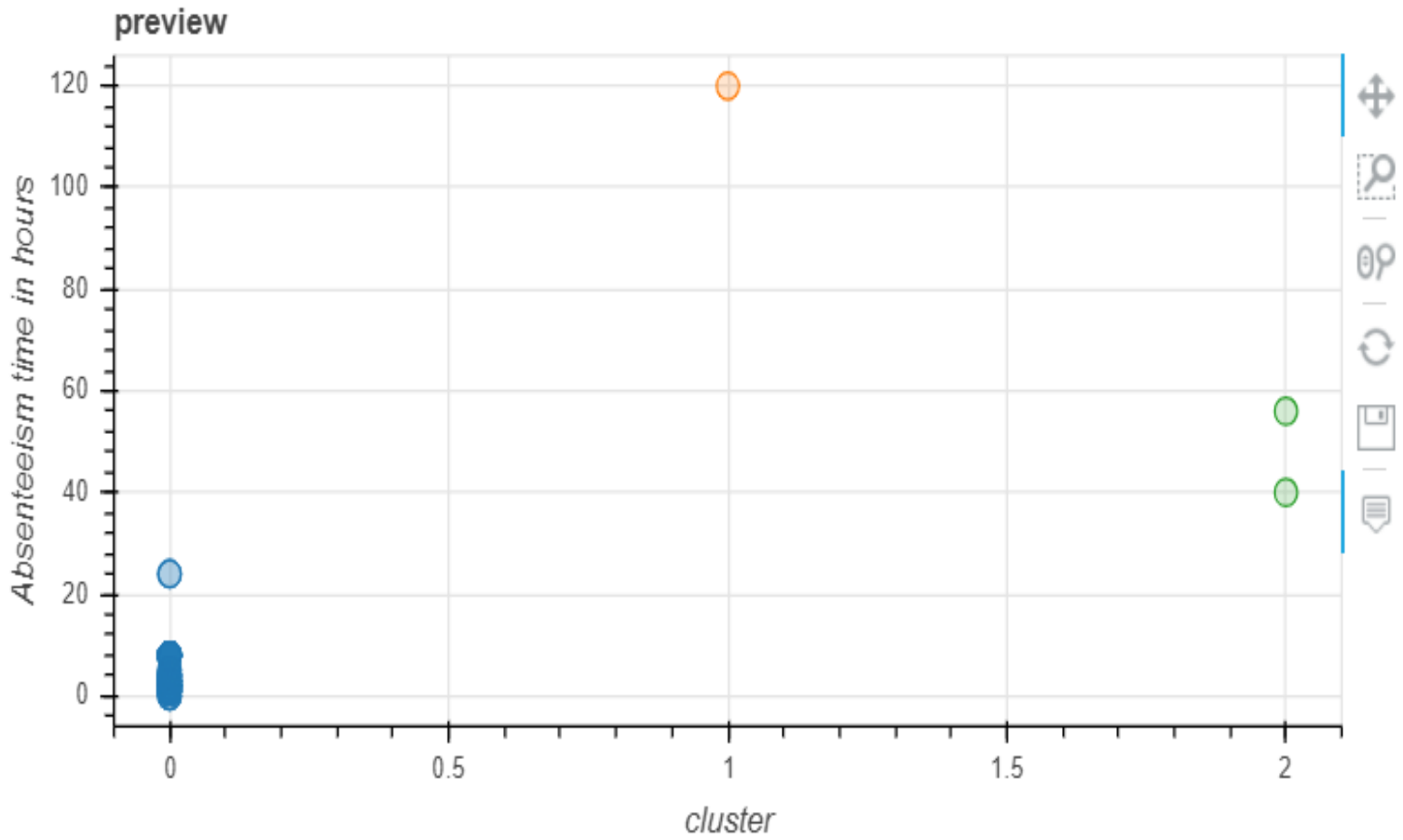
模型

- Agglomerative clustering
- 3 clusters
- linkage=complete
- affinity=euclidean

結果 - Training


preview
Absenteeism time in hours
120
100
80
60
40
20
0
0
0.5
1
1.5
2
cluster

結果 - Testing


preview
Absenteeism time in hours
120
100
80
60
40
20
0
0
0.5
1
1.5
2
cluster

分析

- Absenteeism time in hours 在 40 左右的分類不同
- 資料較少