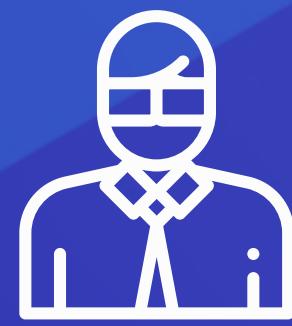




Day 10

機器學習前處理

數值型特徵-去除離群值



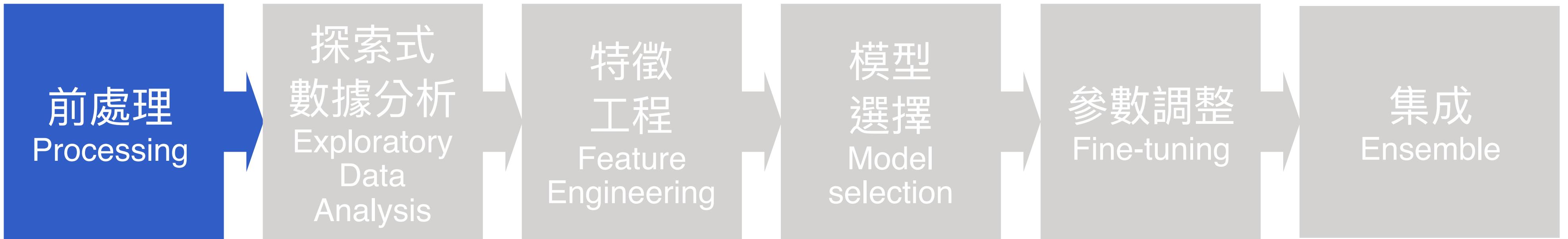
陳明佑

出題教練

知識地圖 數值型特徵 - 去離群值

機器學習前處理

監督式學習 Supervised Learning



非監督式學習 Unsupervised Learning



前處理 Processing



本日知識點目標

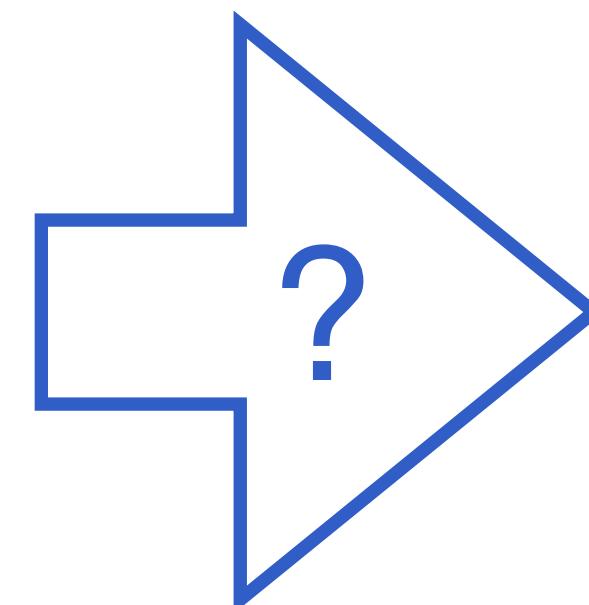
- 初步理解什麼是離群值？出現時會有什麼問題？
- [複習] 異常值處理，會有哪些優缺點？

去除離群值 (1 / 2)

如果只有少數幾筆資料跟其他數值差異很大，標準化無法處理

原始值

50
2
1
0
⋮
2

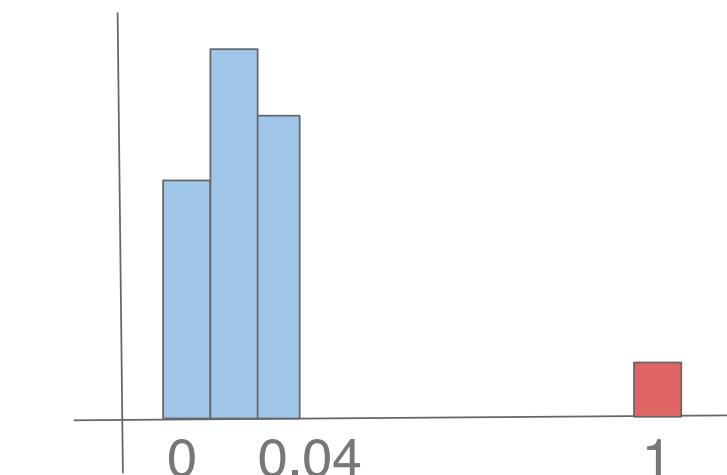
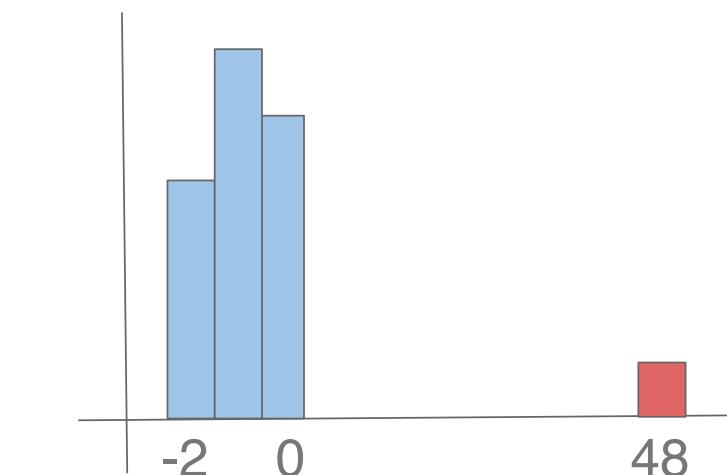
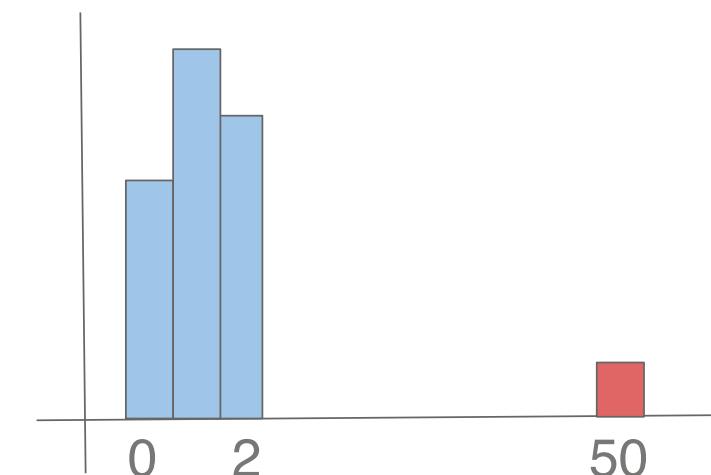


標準化

48
0
-1
-2
⋮
0

最大最小化

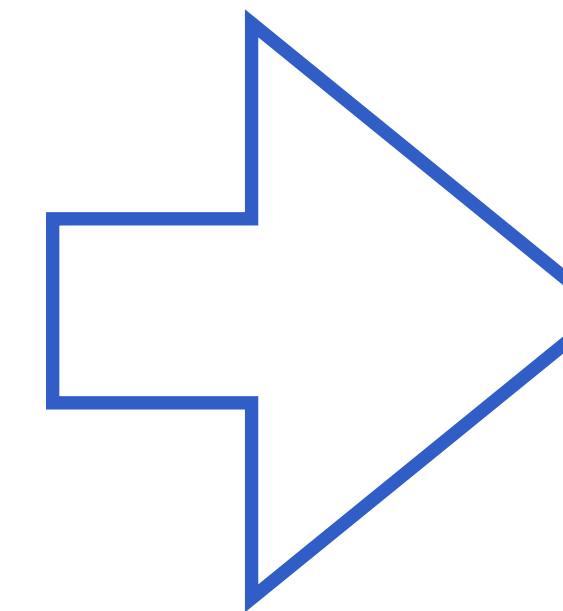
1
0.04
0.02
0
⋮
0.04



去除離群值 (2 / 2)

原始值

50
2
1
⋮
2

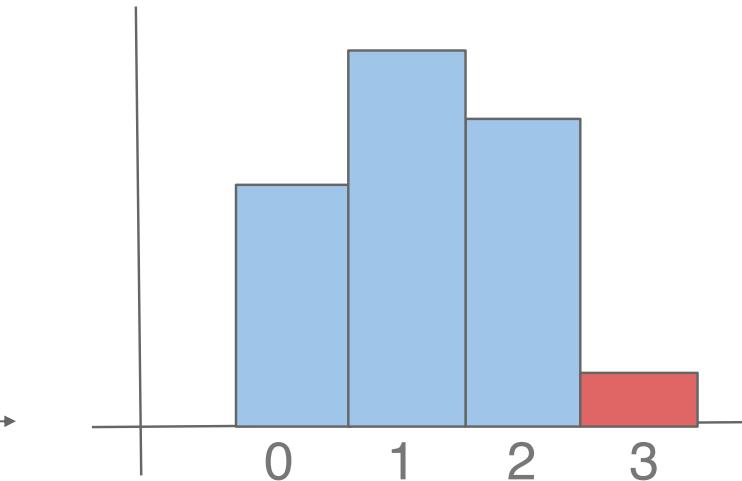
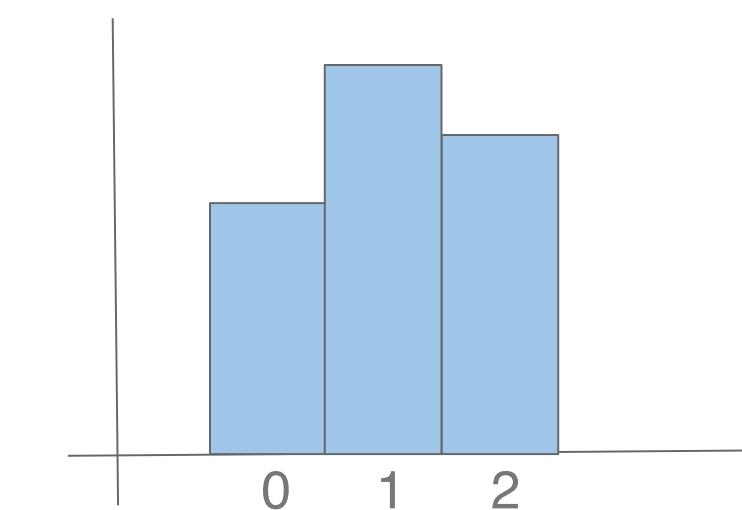
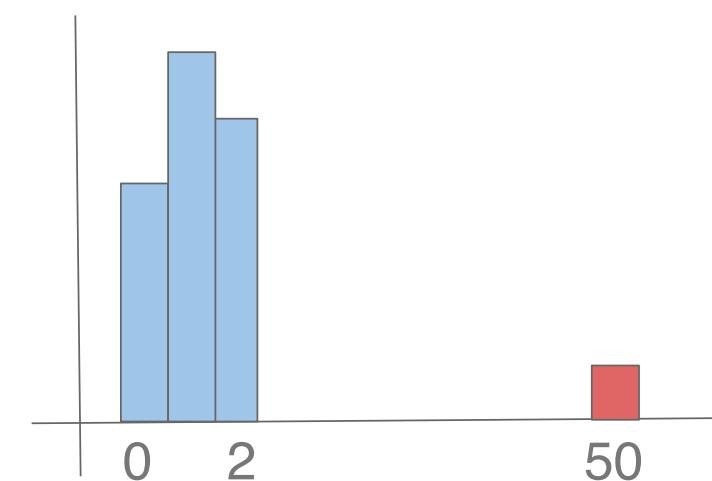


方法一
捨棄離群值

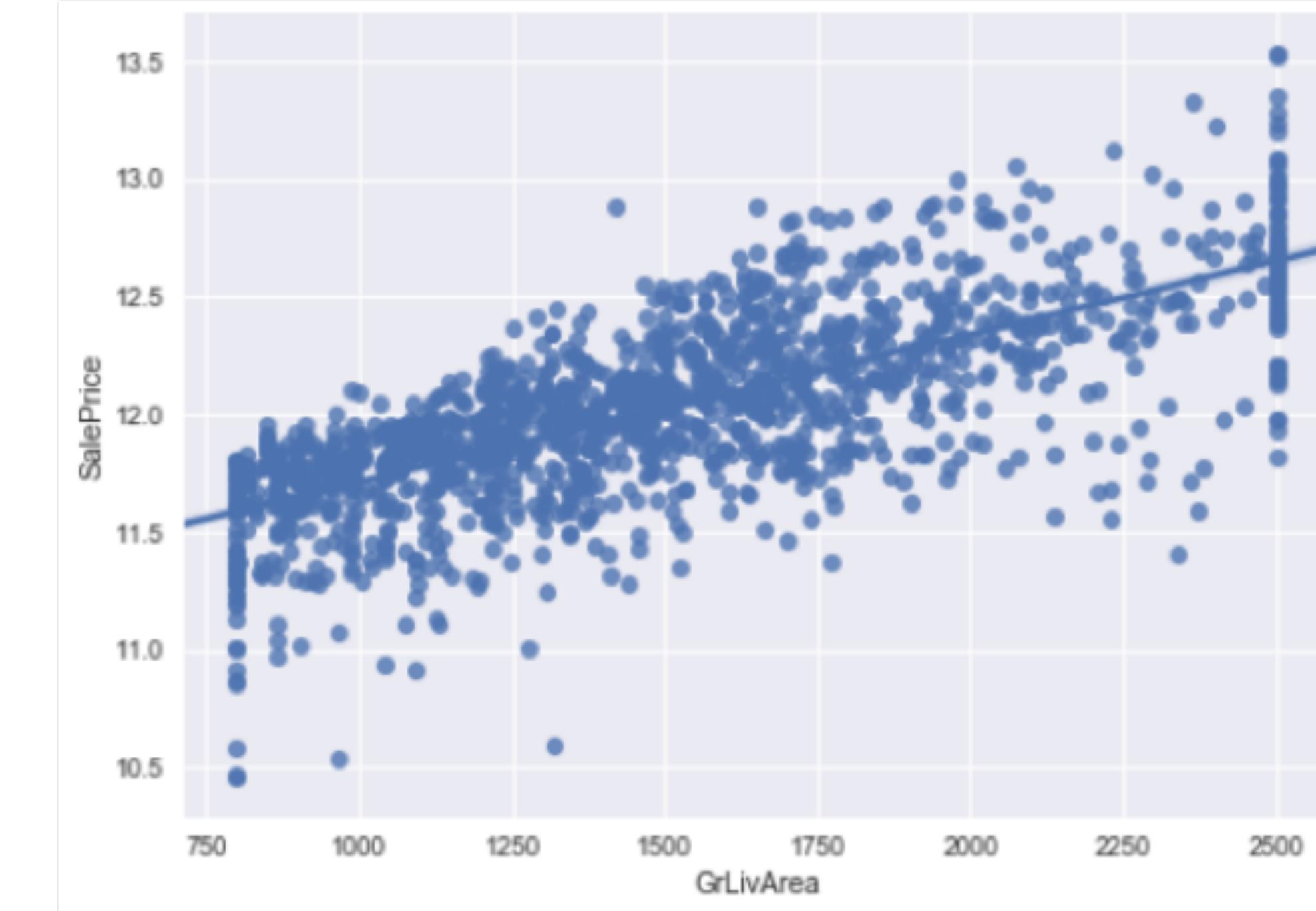
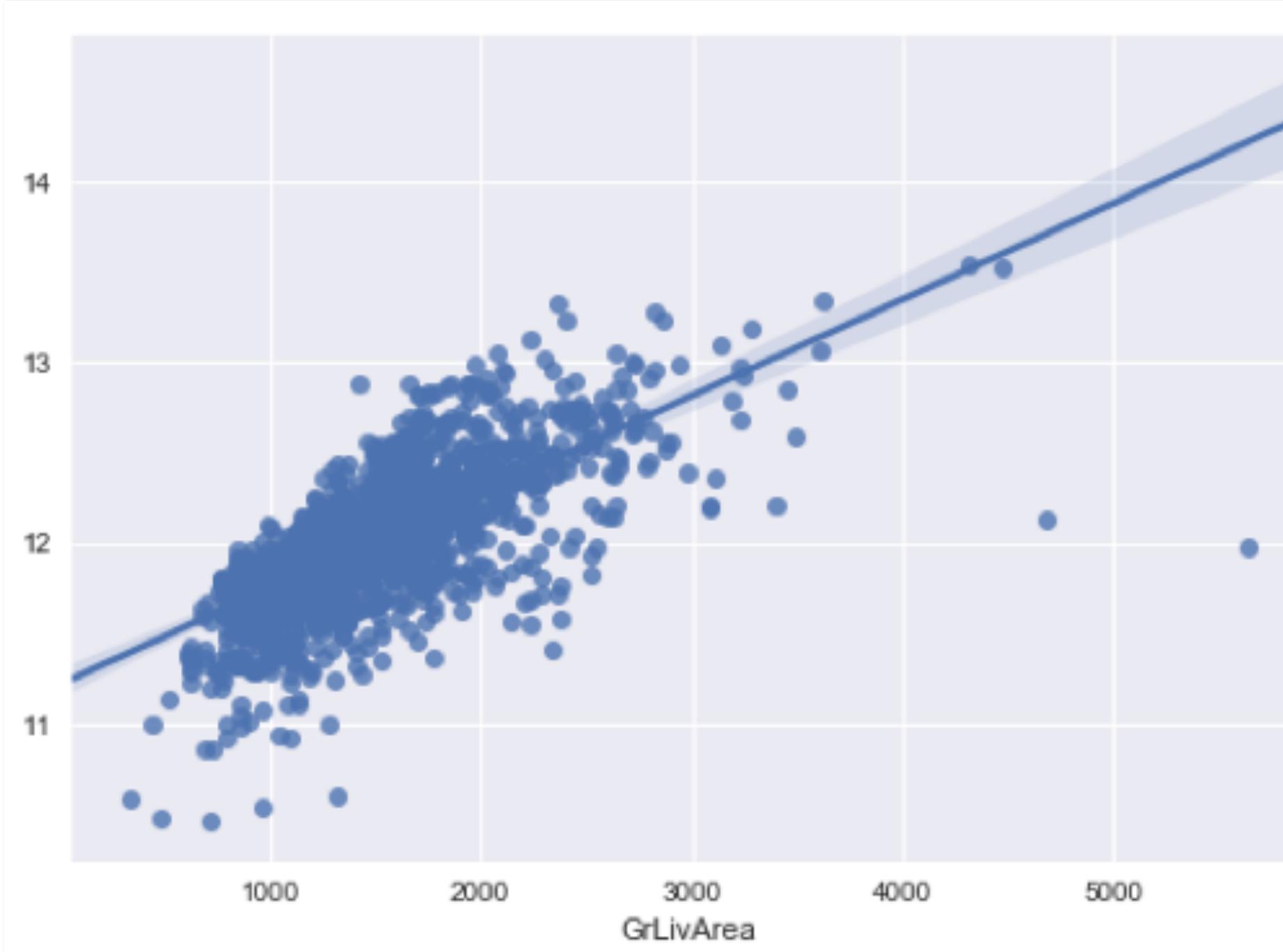
⋮
⋮
2
1
⋮
2

方法二
調整離群值

3
2
1
⋮
2

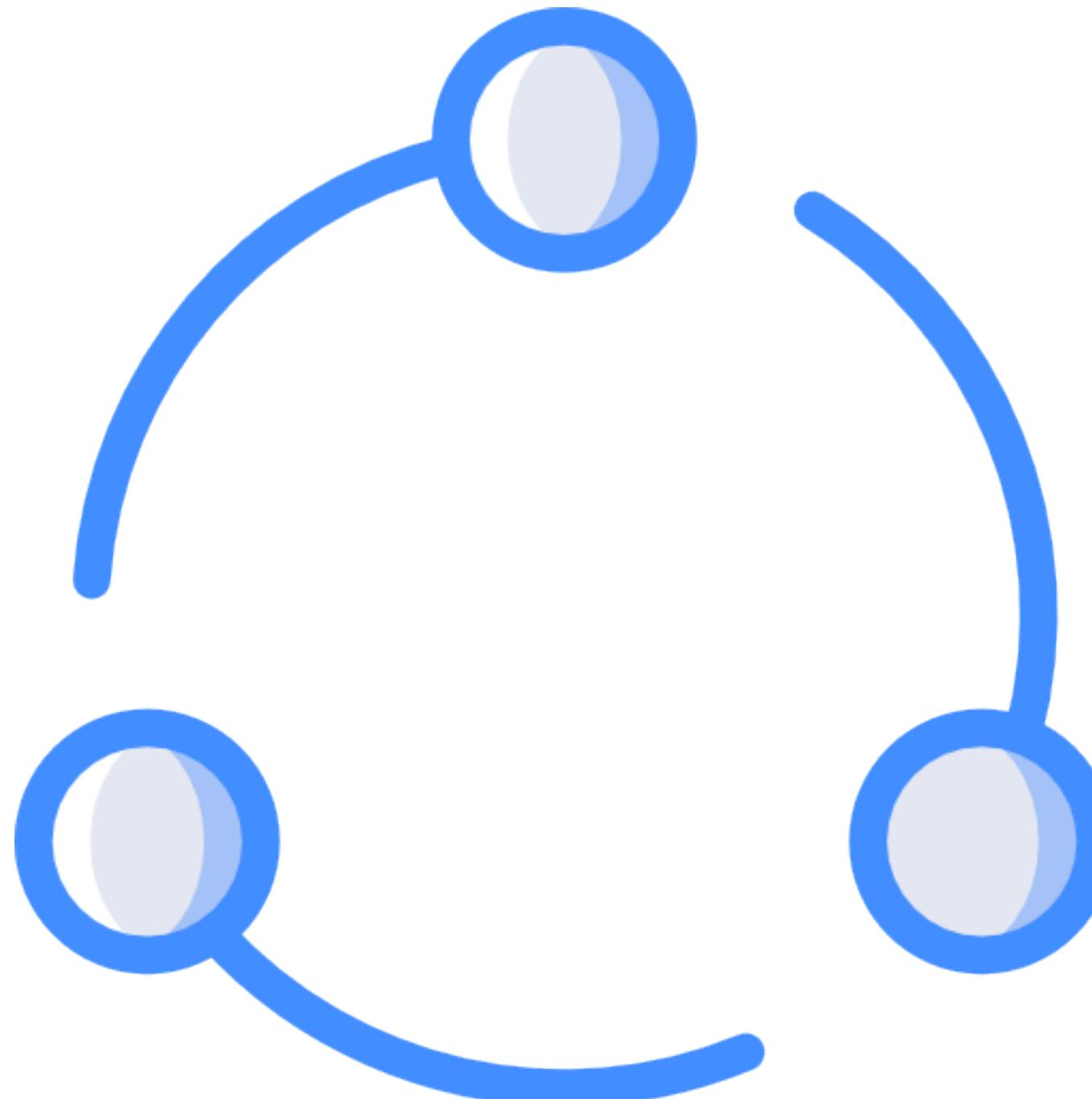


複習：去除離群值



數值型欄位，有時會有與其他數值差距很大的離群值存在
只要離群值數量夠少，除去離群值，將可使得模型預測較為準確
(參考圖中的迴歸直線以及今日範例)

重要知識點複習



- 離群值是與正常數值**偏離較遠**的數值群，如果不處理則
特徵縮放(標準化 / 最小最大化)就會出現很大的問題
- 處理離群值之後，好處是剩餘資料中模型較為**單純且準確**，壞處是有可能**刪除掉重要資訊**，因此刪除前最好能先了解該數值會離群的可能原因



延伸 閱讀

除了每日知識點的基礎之外，推薦的延伸閱讀能補足學員們對該知識點的了解程度，建議您解完每日題目後，若有
多餘時間，可再補充延伸閱讀文章內容。

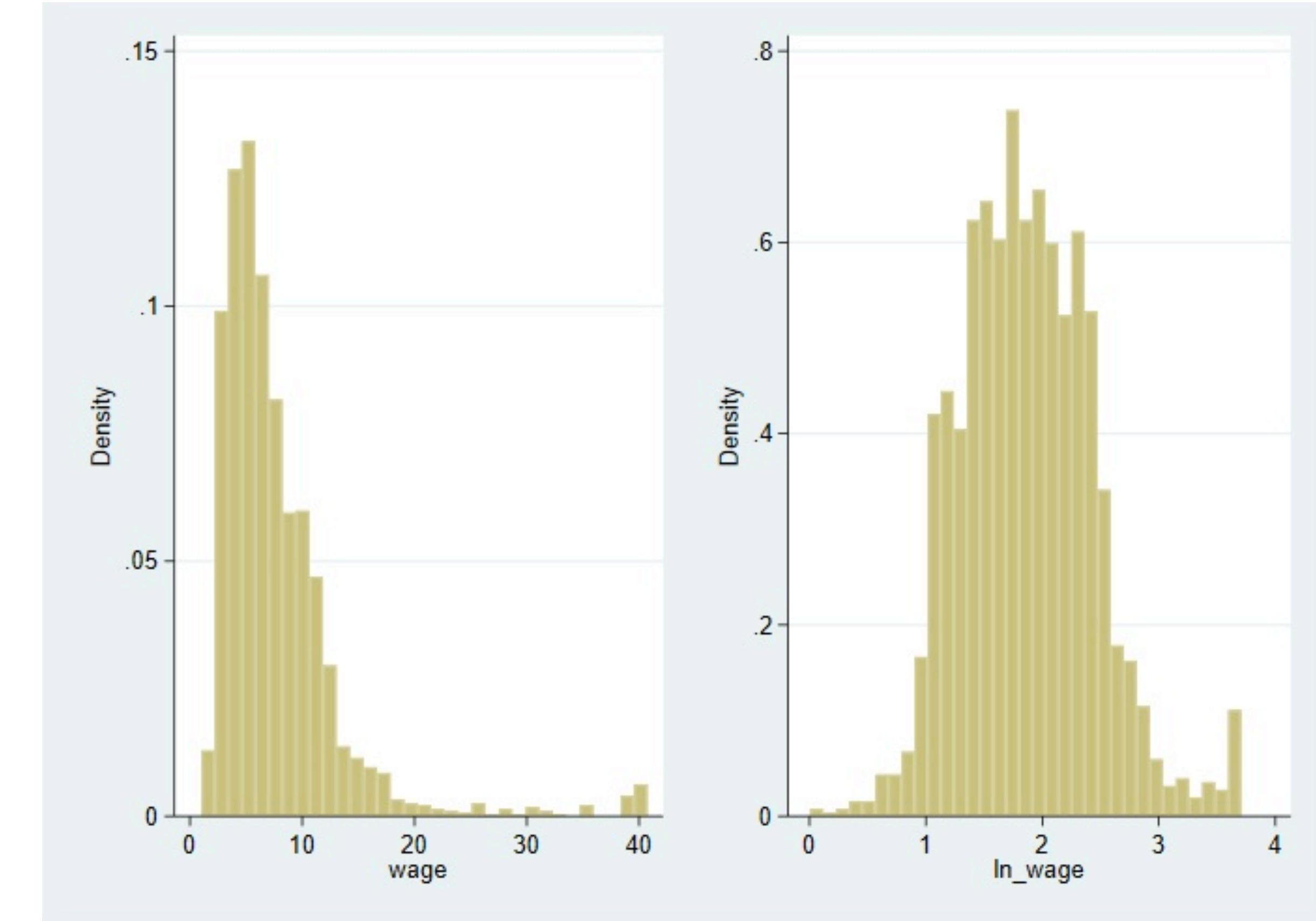
推薦延伸閱讀

離群值! 離群值? 離群值!

Python 知乎-連玉君stata

網頁連結

- 本文除了談到離群值的定義外，主要在第3部分：離群值處理方法，不僅僅告訴你有哪幾類方式，並以圖示的方式呈現讓同學能以常識推論分析，其中對數轉換的部分，我們會在後續的內容有更多的講解，同學大致了解方向即可。





解題時間

It's Your Turn

請跳出PDF至官網Sample Code & 作業
開始解題

