電子商務技術\_HW4

108403201 資管三 A 黃名揚

一. 請用 python 實作以下問題

1. 載入資料並刪除除了"energy", "speechiness", "acousticness", "instrumentalness", "loudness","tempo","danceability",'valence', "liveness" 以外之欄位(使用 pandas dataframe)



2. 將剩下的欄位做特徵篩選的動作，並使用 kmeans silhouette analysis 的方法 找出在哪三個欄位的情況下(需考慮所有組合)，分 X 群會有最高的 silhouette score。請找出 X 與 silhouette score 還有是哪三個欄位。(20%)

請解釋 silhouette 分析法 與 elbow 轉折判斷法的差別(3%)





(欄位:energy、acousticness、loudness)

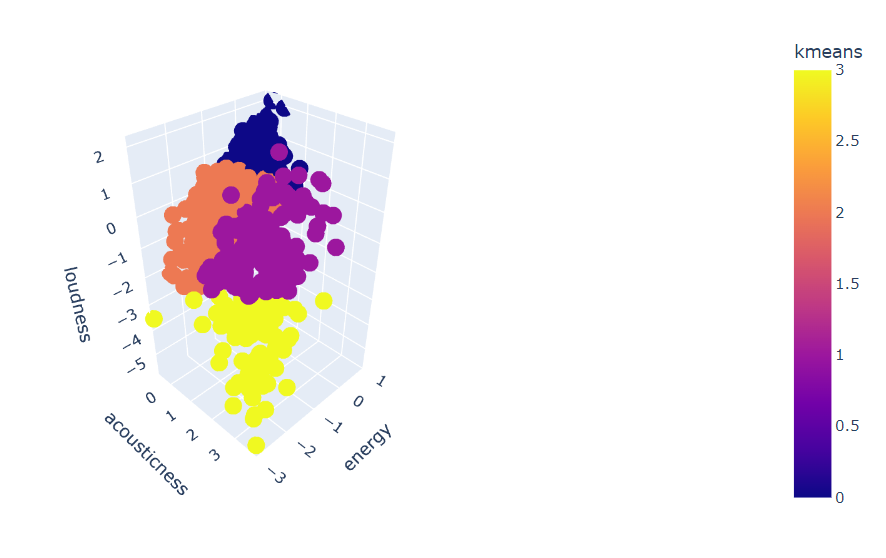
(X=2 silhouette scoure=0.6322586386208744)

silhouette 分析法: 衡量物件和所屬群之間的相似度。Silhouette 值接近1，說明物件與所屬群之間有密切聯繫；反之則接近-1

elbow 轉折判斷法: 適用於K值相對較小的情況。選擇的k值小於真正的時，k每增加1，cost值就會大幅減小；選擇的k值大於真正的K時，k每增加1，cost值的變化不明顯。正確的k值會在轉捩點上。

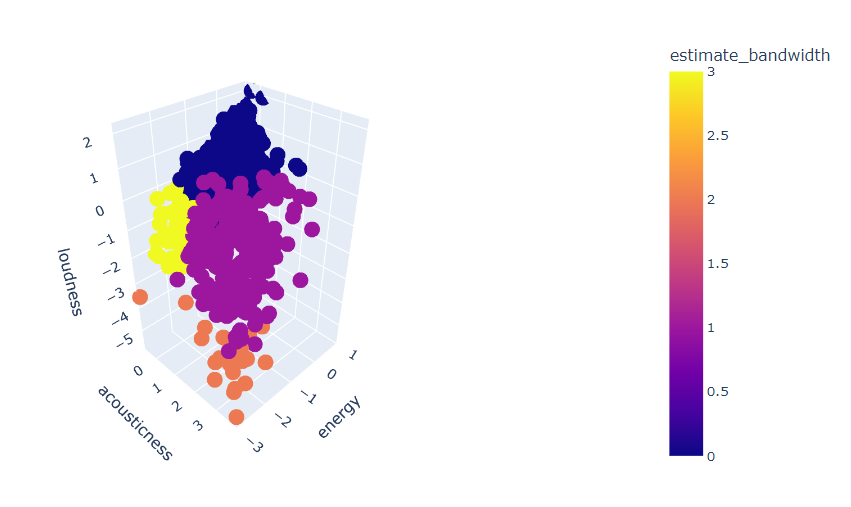
3. 使用剛剛找出來的欄位用 k-means 做分群。超參數設定為 n\_cluster=4 , random\_state=15。 並使用 plotly 繪製出 3d 圖形如以下所示(15%)： 注意要有欄位名稱，也就是剛剛找出來的那三個

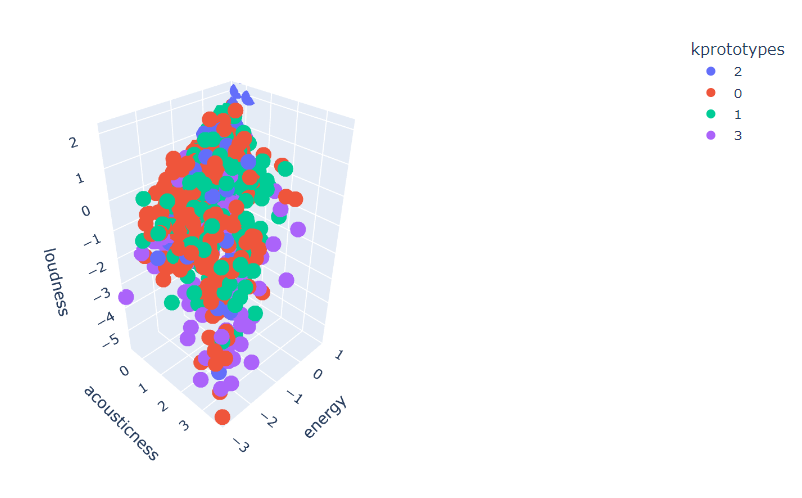




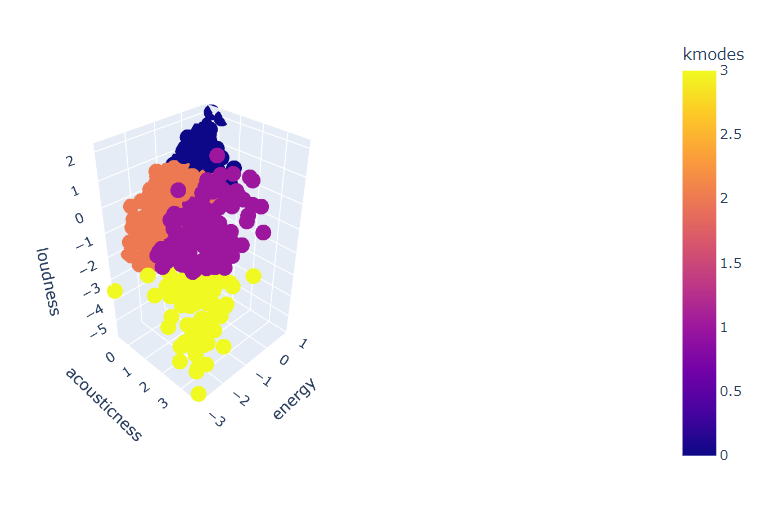
4. 使用剛剛找出來的欄位用 Meanshift 做分群(15%) 請找出最佳的 estimate\_bandwidth.超參數設定為 random\_state=15, quantile=0.32, n\_samples=1000 使用剛剛找出的 estimate\_bandwidth 做分群並繪製如第三題的圖





5. 使用剛剛找出來的欄位用 k-prototypes 做分群。超參數設定為 n\_cluster=4 random\_state=15,init=’Huang’,verbose=0。 並使用 plotly 繪製出 3d 圖形如第三題的圖

6. 使用剛剛找出來的欄位用 k-modes 做分群。超參數設定為 n\_cluster=4 random\_state=15,init=’Huang’,verbose=0。 並使用 plotly 繪製出 3d 圖形如第三題的圖



7. 請比較說明上述四種分群法的差異

k-means:適用於數字資料

k-modes:適用於類別資料

k-prototype: k-means與k-modes的結合，可用於數字與類別的混合資料

Meanshift: 對類別個數未知的數據，k-means較難求出解。基於密度的非參數聚類算法，但與k-means不同的地方在於不須事先制定k值

二. 使用 weka 做分群

