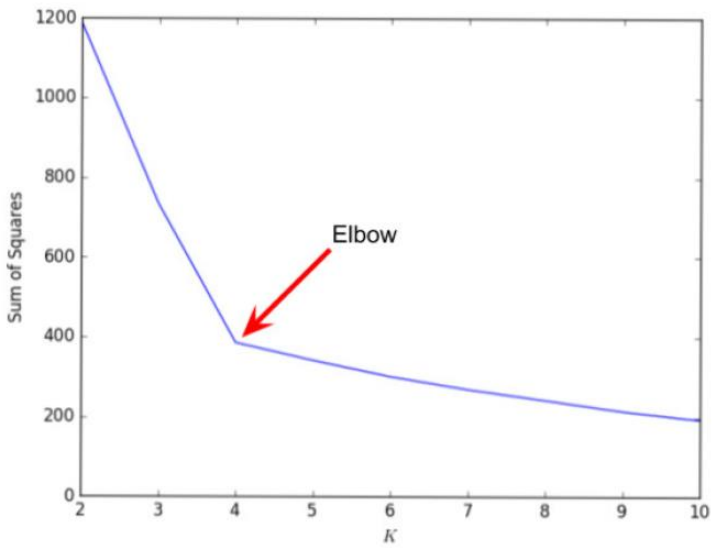
2021電子商務技術作業五

請使用 python 回答以下問題，截圖程式碼與執行結果加上說明文字：

使用 ageinc.csv 資料集，使用 k-means 分群法進行分群

1. 將 income 與 age 標準化，建立成標準化後的新欄位。 (2%)
2. 使用 Elbow 方法找出最佳的分群數量 k：

   Elbow 方法：找尋最佳 k 值使 SSE (sum of squared errors，即資料點與分群中心點的距離平方和) 最小化。當選擇的 k 值小於最佳 k 值時，k 每增加 1，SSE 就會大幅的減小；當選擇的 k 值大於最佳 k 值時，k 每增加 1，SSE 的變化不會那麼明顯。這樣，最佳 k 值就會在這個轉捩點，類似 elbow 的地方，如下圖：



$$SS = \sum_{k} \sum_{x_i \epsilon k} (x_i - \mu_k)^2$$

   (1) 計算出 k-means 分群數量為 2 時的 SSE 值，random_state 設 10。(5%)

   (2) 使用迴圈計算出 k-means 分群數量為 2 至 10 的 SSE 值，random_state 設 10。(5%)

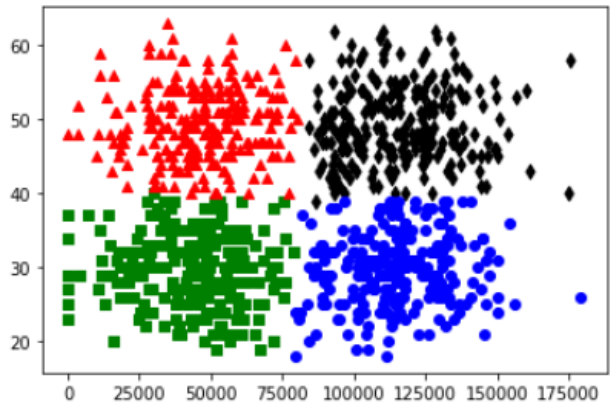
   (3) 畫出 Elbow 圖 (如上方說明圖)，說明最佳分群數為幾個。(10%)

3. 使用最佳 k 進行 k-means 分群，random_state 設 10，將分群結果加在 DataFrame 欄位中並列印出來，如下圖所示。(10%)

	income	age	z_income	z_age	kmean_cluster
0	101743	58	0.550812	1.693570	0
1	49597	27	-0.777331	-1.130565	1
2	36517	52	-1.110474	1.146963	2
3	33223	49	-1.194372	0.873660	2
4	72994	53	-0.181416	1.238064	2
...	...	...	...	...	...
995	70615	29	-0.242008	-0.948363	1
996	95102	41	0.381668	0.144851	0
997	42203	35	-0.965654	-0.401756	1
998	16975	31	-1.608203	-0.766161	1
999	123857	44	1.114049	0.418154	0

使用 ageinc.csv 資料集，使用 mean-shift 分群法進行分群

4. 將 income 與 age 標準化，建立成標準化後的新欄位。(1%)
5. 用 estimate_bandwidth 找出最佳帶寬，quantile 設 0.1。(5%)
6. 用 mean-shift 分群，將分群結果加在 DataFrame 欄位中並列印出來，並說明分群數。(10%)
7. 使用 matplotlib 繪出如下圖的 k-means 與 mean-shift 分群結果 (x軸為 income，y軸為age)。(10%)



使用 age_education.csv 資料集，使用 k-prototypes 分群法進行分群

8. 將 age 標準化，建立成標準化後的新欄位。(1%)
9. 用 k-prototypes 分群，分群數量設 3，random_state 設 10，並印出每個分群中的資料數量。(10%)

使用 customer_offer.csv 資料集，使用 k-modes 分群法進行分群

10. 將 customer_name 設為 index 以利後續分群。(2%)

11. 用 k-modes 分群，分群數量設 4，random_state 設 10，並印出每個分群中的資料數量。(9%)
12. 請用文字說明 k-means、mean-shift、k-prototypes、k-modes 四種分群方法的差異。(10%)

請使用 WEKA 回答以下問題，截圖操作過程與執行結果加上說明文字：

13. 使用 simpleKMeans 對 ageinc.csv 進行分群，分群數量設定 4，並與 python 的執行結果做比較。(10%)

- **繳交期限：4/21(三)中午12:00**
- **請繳交 .pdf檔和 .ipynb 檔，檔名為ECT_HW5_學號**
- **上傳至ee-class 作業區，遲交一天扣該次作業5%**