課題 1

cs3-06-assign1

データ market.csv, ID_data.csv はそれぞれ、ある小売店の30日間の売り上げおよび顧客属性のデータである。これらを用いて、以下の操作を行うスクリプトをJupyterで作成せよ(ノートブック名は cs3-06-assign1.ipynb/html とせよ)

1. 必要なライブラリをimport。
2. market.csv, ID_data.csv のデータをデータフレーム df_market, df_id にそれぞれ読み込み、行数と列数、各列のデータ型と欠損値でないデータの数、先頭5行と末尾5行を表示して確認。
3. df_market と df_id を、「顧客ID」列をキーとして左外結合で結合し、df に代入。dfの先頭5行を表示して確認。
4. dfから、「年代」列の値が50のデータのみを抜き出し、df_50に代入。この時、indexを0からの連番に振り直すこと。また元のindexは不要。
5. df_50から「個数」列と「税抜価格」列のみを抜き出したデータフレーム dfX を作成する。行数と列数、先頭5行を表示して確認。
6. dfXの各列を平均0、母標準偏差(偏差二乗和をデータ数Nで割った分散の平方根)1に標準化し、変数 X_scaled に代入。さらにX_scaled の平均と母標準偏差を表示。

7. X_scaled のデータ型と、行数・列数を表示。
8. X_scaled に dfX と同じ列ラベルを付与したデータフレーム dfX_scaled を作成し、データ型を確認、さらに先頭5行を表示。
9. dfX_scaled に対して、KMeans法 (n_init=10とする) によるElbow法を、最大クラスタ数10で実施。クラスタ数を横軸、Inertiaを縦軸とするグラフを描画。
10. dfX_scaled に対してクラスタ数4で KMeans法によるクラスタリングを実行。このとき、n_init=10, random_state=5 とせよ。クラスタリング結果を変数 cls に代入、表示して確認。先頭のデータが割り当てられたクラスタ番号を答えよ。
11. データフレームdf_50に、新たな列「cluster_no」を追加し、各データが属するクラスタの番号を格納する。先頭5行を表示して確認。
12. 「cluster_no」列の各値の出現数(各クラスタのメンバー数) を表示。
13. df_50の「個数」、「税抜価格」をそれぞれ横軸、縦軸として散布図を描画。クラスタごとに色をつけて区別できるようにせよ。