**K mean giá nhà**

# -----------------------------

# 1️⃣ Import thư viện

# -----------------------------

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.datasets import fetch\_california\_housing

from sklearn.preprocessing import StandardScaler

from sklearn.cluster import KMeans

import seaborn as sns

# -----------------------------

# 2️⃣ Load dữ liệu giá nhà

# -----------------------------

housing = fetch\_california\_housing()

X = pd.DataFrame(housing.data, columns=housing.feature\_names)

y = pd.Series(housing.target, name='MedHouseVal')

# -----------------------------

# 3️⃣ Chuẩn hóa dữ liệu

# -----------------------------

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

# -----------------------------

# 4️⃣ Huấn luyện K-Means

# -----------------------------

k = 5 # số cluster

kmeans = KMeans(n\_clusters=k, random\_state=1)

clusters = kmeans.fit\_predict(X\_scaled)

# Thêm cột cluster vào DataFrame

X['Cluster'] = clusters

X['MedHouseVal'] = y

# -----------------------------

# 5️⃣ Thống kê giá trung bình theo cluster

# -----------------------------

cluster\_summary = X.groupby('Cluster')['MedHouseVal'].agg(['mean','median','count'])

print(cluster\_summary)

# -----------------------------

# 6️⃣ Biểu diễn trực quan 2D (theo 2 feature chính)

# -----------------------------

plt.figure(figsize=(8,6))

sns.scatterplot(x=X['MedInc'], y=X['MedHouseVal'], hue=clusters, palette='Set1', alpha=0.6)

plt.xlabel('MedInc (Median Income)')

plt.ylabel('MedHouseVal (Median House Value)')

plt.title('K-Means Clustering of Housing Data')

plt.legend(title='Cluster')

plt.show()

**HAC giá nhà**

# -----------------------------

# 1️⃣ Import thư viện

# -----------------------------

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.datasets import fetch\_california\_housing

from sklearn.preprocessing import StandardScaler

from scipy.cluster.hierarchy import linkage, dendrogram, fcluster

# -----------------------------

# 2️⃣ Load dữ liệu giá nhà

# -----------------------------

housing = fetch\_california\_housing()

X = pd.DataFrame(housing.data, columns=housing.feature\_names)

y = pd.Series(housing.target, name='MedHouseVal')

# -----------------------------

# 3️⃣ Chuẩn hóa dữ liệu

# -----------------------------

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

# -----------------------------

# 4️⃣ Tính ma trận liên kết (linkage matrix)

# -----------------------------

Z = linkage(X\_scaled, method='ward') # ward, complete, average, single

# -----------------------------

# 5️⃣ Vẽ dendrogram

# -----------------------------

plt.figure(figsize=(12,6))

dendrogram(Z, truncate\_mode='level', p=5, leaf\_rotation=45., leaf\_font\_size=12.)

plt.title('Hierarchical Agglomerative Clustering Dendrogram')

plt.xlabel('Samples')

plt.ylabel('Distance')

plt.show()

# -----------------------------

# 6️⃣ Lấy cluster từ dendrogram

# -----------------------------

k = 5 # số cluster mong muốn

clusters = fcluster(Z, k, criterion='maxclust')

# Thêm cluster vào DataFrame

X['Cluster'] = clusters

X['MedHouseVal'] = y

# -----------------------------

# 7️⃣ Thống kê giá trung bình theo cluster

# -----------------------------

cluster\_summary = X.groupby('Cluster')['MedHouseVal'].agg(['mean','median','count'])

print(cluster\_summary)

# -----------------------------

# 8️⃣ Biểu diễn trực quan 2D (theo 2 feature chính)

# -----------------------------

plt.figure(figsize=(8,6))

plt.scatter(X['MedInc'], X['MedHouseVal'], c=clusters, cmap='Set1', alpha=0.6)

plt.xlabel('MedInc (Median Income)')

plt.ylabel('MedHouseVal (Median House Value)')

plt.title('HAC Clustering of Housing Data')

plt.show()