Національний технічний університет України «КПІ»

Кафедра автоматизованих систем обробки інформації і управління

Лабораторна робота №6

з дисципліни «[Аналіз даних в інформаційних управляючих системах](https://do.ipo.kpi.ua/course/view.php?id=5470)»

на тему: «НЕІЄРАРХІЧНІ МЕТОДИ КЛАСТЕРНОГО АНАЛІЗУ»

Виконав:

студент групи ІС-23

Шимків М.В.

Викладач:

Гавриленко О.В

Київ 2024

**Мета роботи:** Практичне засвоєння неієрархічного кластерного аналізу

багатовимірних даних на приклады методу k-середніх.

**Завдання до роботи**

1. Ознайомитися з конспектом лекцій та рекомендованою

літературою, а також додатком Е, що містить короткі теоретичні відомості

про неієрархічні методи кластерного аналізу та особливості їх застосування

в Matlab.

2. Вивчити функції MATLAB: scatter, gscatter, min, pdist, hist, std.

3. Завантажити дані відповідно до вашого варіанту (табл. 1).

Побудувати графічне зображення експериментальних даних (діаграму

розсіювання). Візуально оцінити кількість кластерів k за побудованим

зображенням.

4. Розробити алгоритм кластеризації k-середніх і програмно його

реалізувати в середовищі MATLAB.

5. Виконати кластерний аналіз висхідних даних методом k-середніх

(параметри методу див. в табл. 2). Визначити найбільш оптимальну

кількість кластерів за допомогою методу «Ліктя» або методу «Силуету».

6. Розрахувати центри отриманих кластерів. Відобразити графічно

знайдені кластери (використати діаграму розсіювання у кольорі).

7. Оформити звіт з роботи.

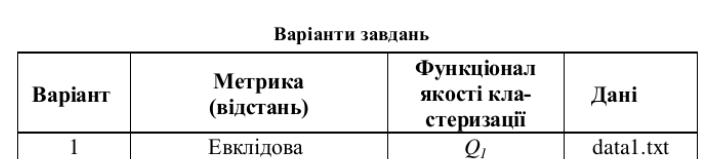
8. Відповісти на контрольні питання.

**Набір даних:**

Набір даних я взяв з сервісу kaggle. Ось посилання на набір даних:

<https://www.kaggle.com/datasets/varunraskar/store-sales-data>

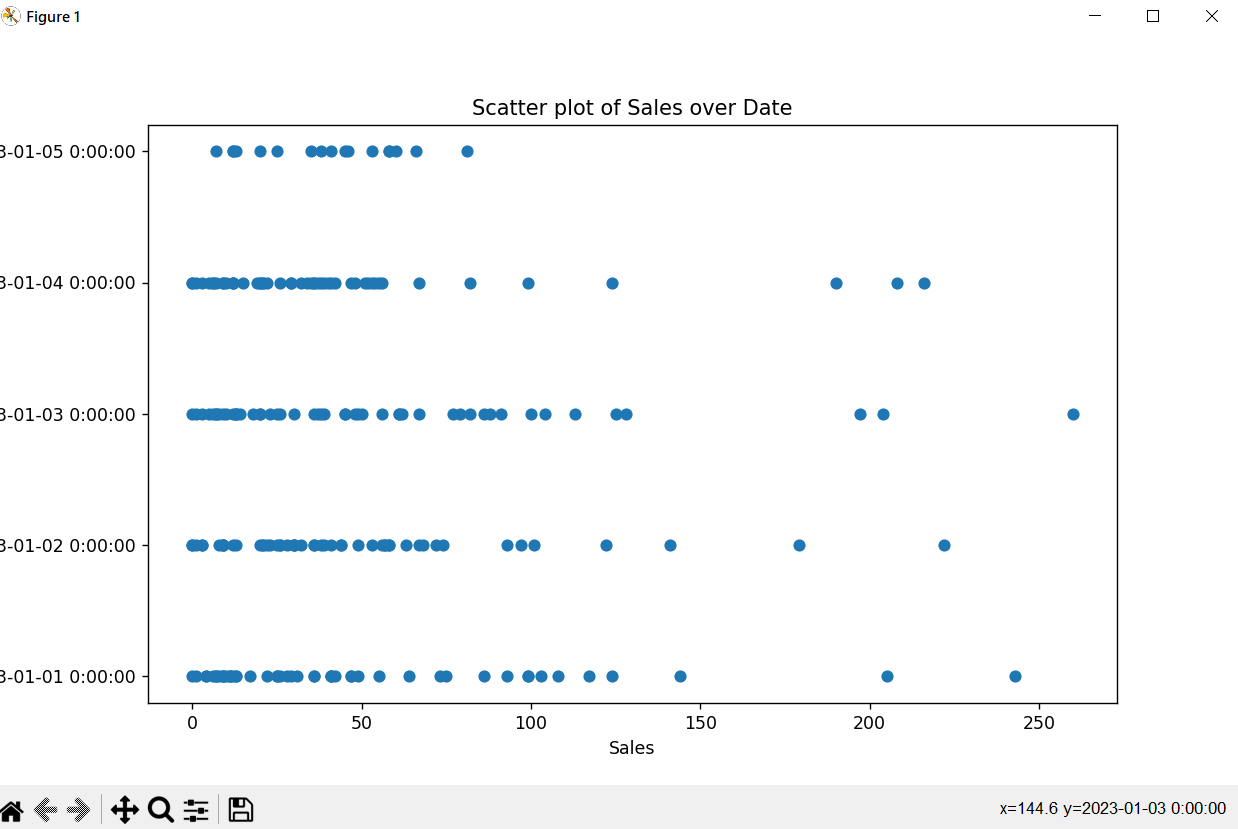
**Варіант 25**

****

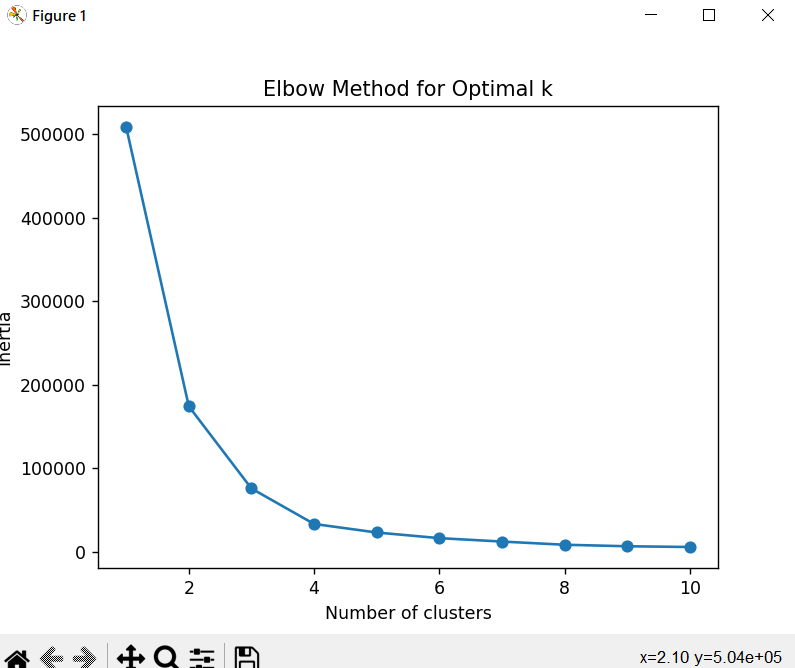
Запуск програми відбувається так:



Відображення діаграми розсіювання:



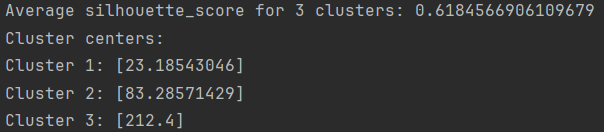
Цей графік відображає результати використання методу ліктя (Elbow Method) для визначення оптимальної кількості кластерів у моделі k-means кластеризації. Інерція (Inertia) є мірою внутрішньокластерної варіабельності в кластерах:



Відображення графічного представлення кластерів:



Оцінка силуета та розрахунок центрів кластерів:



**Код програми:**

import pandas as pd  
import matplotlib.pyplot as plt  
from sklearn.cluster import KMeans  
from sklearn.metrics import silhouette\_score  
  
# Завантаження даних з CSV  
data = pd.read\_csv('Sales2.csv')  
  
# Відображення діаграми розсіювання  
plt.figure(figsize=(10, 6))  
plt.scatter(data['Sales'], data['Date'])  
plt.xlabel('Sales')  
plt.ylabel('Date')  
plt.title('Scatter plot of Sales over Date')  
plt.show()  
  
# Оцінка кількості кластерів k  
# На графіку розсіювання можна помітити відносно ясно визначені групи,  
# так що можна спробувати використати метод "Ліктя" для вибору оптимальної кількості кластерів.  
inertia = []  
for k in range(1, 11):  
 kmeans = KMeans(n\_clusters=k, random\_state=42)  
 kmeans.fit(data[['Sales']])  
 inertia.append(kmeans.inertia\_)  
  
plt.plot(range(1, 11), inertia, marker='o')  
plt.xlabel('Number of clusters')  
plt.ylabel('Inertia')  
plt.title('Elbow Method for Optimal k')  
plt.show()  
  
# Використання методу "Ліктя" для визначення оптимальної кількості кластерів  
# Зазвичай, "ліктьова точка" є місцем, де зміна відносної дії зменшується значно.  
# У цьому випадку можна вибрати k=3.  
optimal\_k = 3  
  
# Кластерний аналіз методом k-середніх  
kmeans = KMeans(n\_clusters=optimal\_k, random\_state=42)  
data['Cluster'] = kmeans.fit\_predict(data[['Sales']])  
  
# Оцінка силуета  
silhouette\_avg = silhouette\_score(data[['Sales']], data['Cluster'])  
print(f"Average silhouette\_score for {optimal\_k} clusters:", silhouette\_avg)  
  
# Розрахунок центрів кластерів  
cluster\_centers = kmeans.cluster\_centers\_  
print("Cluster centers:")  
for i, center in enumerate(cluster\_centers):  
 print(f"Cluster {i+1}: {center}")  
  
# Відображення графічного представлення кластерів  
plt.figure(figsize=(10, 6))  
colors = ['blue', 'green', 'red']  
for i in range(optimal\_k):  
 cluster\_data = data[data['Cluster'] == i]  
 plt.scatter(cluster\_data['Sales'], cluster\_data['Date'], color=colors[i], label=f'Cluster {i+1}')  
  
plt.scatter(cluster\_centers, [data['Date'].min()] \* optimal\_k, color='black', marker='x', label='Cluster Centers')  
plt.xlabel('Sales')  
plt.ylabel('Date')  
plt.title('Clustered Sales over Date')  
plt.legend()  
plt.show()

**Контрольні питання:**

1. Задача неієрархічного кластерного аналізу полягає у розбитті множини об'єктів на групи таким чином, щоб об'єкти всередині одного кластера були подібні між собою, а об'єкти з різних кластерів були відмінні.
2. Методи неієрархічного кластерного аналізу використовуються для різноманітних задач обробки експериментальних даних, таких як сегментація клієнтів у маркетингових дослідженнях, аналіз текстів для категоризації документів, класифікація зображень у комп'ютерному зорі та багато інших.
3. Сутність алгоритму k-середніх полягає в обранні початкових центрів кластерів, призначенні кожного об'єкта до найближчого центру кластера, оновленні центрів кластерів і повторенні цих кроків до збіжності.
4. Основні етапи алгоритму k-середніх включають ініціалізацію початкових центрів кластерів, призначення кожного об'єкта до найближчого центру кластера, оновлення центрів кластерів та повторення призначення та оновлення до збіжності.
5. Матриця розбиттів U вводиться в алгоритмі k-середніх для відстеження належності кожного об'єкта до певного кластера. Кожен елемент матриці представляє ймовірність того, що об'єкт належить до певного кластера.
6. На практиці для автоматичної кластеризації можуть використовуватися такі критерії зупинки, як досягнення максимальної кількості ітерацій, досягнення сталої інерції або досягнення збіжності.
7. Кількість кластерів в алгоритмі k-середніх можна оцінити за допомогою методу ліктя (Elbow Method), який використовується для вибору оптимальної кількості кластерів, спостерігаючи за зміною інерції відносно кількості кластерів.
8. Функції MATLAB, такі як scatter, gscatter, min, pdist, hist, std, використовуються для візуалізації даних, обчислення статистичних характеристик, обробки відстаней між об'єктами та інших завдань, пов'язаних з кластеризацією та обробкою даних.

**Висновок:**

Отже, в результаті проведення лабораторної роботи з неієрархічного кластерного аналізу було виявлено, що оптимальною кількістю кластерів для наших даних є три. За допомогою алгоритму k-середніх були знайдені центри кожного кластера, а відповідність об'єктів до них візуалізовано на діаграмі розсіювання. Отримані результати дозволяють зробити висновок про наявність трьох унікальних груп даних з відмінними характеристиками.