Національний технічний університет України «КПІ ім. Ігоря Сікорського» Факультет Інформатики та Обчислювальної Техніки

Кафедра інформаційних систем та технологій

Лабораторна робота №5

з дисципліни «Інтелектуальний Аналіз Даних»

на тему

«ІЄРАРХІЧНІ МЕТОДИ КЛАСТЕРНОГО АНАЛІЗУ»

Варіант №12

Виконав:

студент групи ІС-02

Плостак І. М.

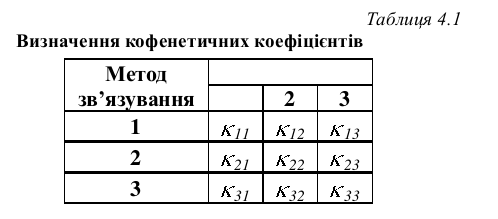
Київ – 2021

**1. Мета роботи**

Практичне засвоєння методів ієрархічного кластерного аналізу даних.

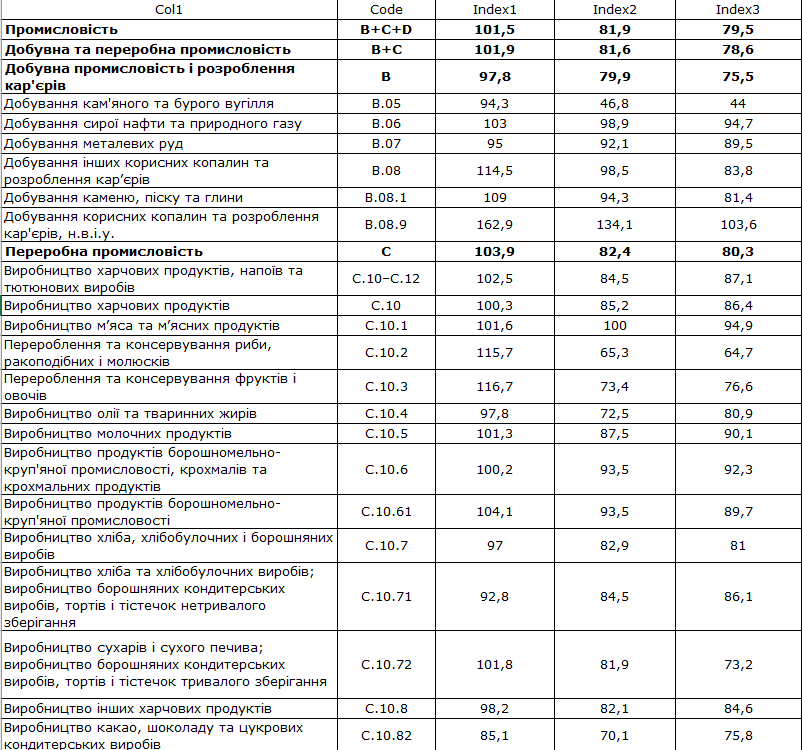
**2. Завдання до роботи**

1. Ознайомитися з конспектом лекцій та рекомендованою літературою, а також додатком Д, що містить короткі теоретичні відомості про ієрархічні методи кластерного аналізу та особливості їх застосування.
2. За допомогою команди >>help вивчити функції pdist, squareform, linkage, dendrogram, cophenet, cluster, gscatter.
3. Завантажити висхідний набір даних для обробки та аналізу згідно до вашого варіанту (табл. 4.2). Побудувати графічне зображення експериментальних даних.
4. Обчислити відстань між об'єктами. Використовувати міри для розрахунку відстаней відповідно до варіанту (табл. 4.2).
5. Використовуючи рекомендоване програмне забезпечення здійснити кластерний аналіз висхідних даних методом ієрархічної кластеризації за методами зв'язування відповідно до варіанту (табл. 4.2).
6. Виконати аналіз якості кластеризації за допомогою обчислення кофенетичного кореляційного коефіцієнту. Заповнити таблицю для кофенетичного кореляційного коефіцієнту (табл. 4.1).



1. Визначити найбільш і найменш ефективні способи ієрархічної кластеризації для аналізу висхідного набору даних (максимальні і мінімальні коефіцієнти та відповідні до них способи кластеризації). Для найбільш ефективного способу ієрархічної кластериації побудувати дендрограму результатів кластерного аналізу.
2. Визначити кількість вірогідних кластерів. Для виділення значущих кластерів використовувати порогове значення, розраховане за метрикою відстаней або методом завдання фіксованого числа кластерів.
3. Розрахувати центри та внутрішньокластерну дисперсію отриманих кластерів, геометричні відстані від елементів до центрів кластерів, відстані між центрами кластерів. Відобразити графічно знайдені кластери та ії центри (скористатися діаграмою розсіювання у кольорі).
4. Оформити звіт з роботи.
5. Відповісти на контрольні питання.

**3. Набір даних для обробки (частина)**



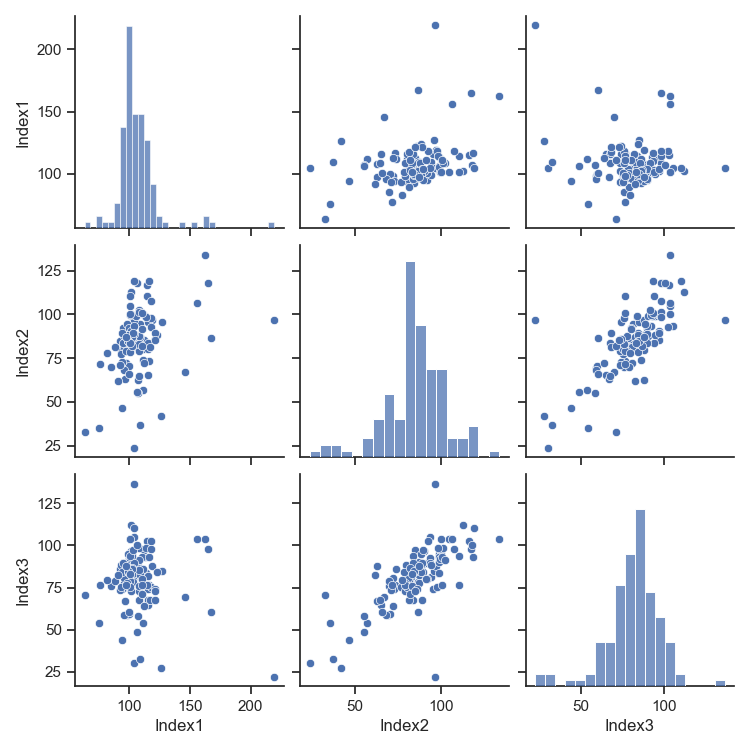
**4. Лістинг основних функцій програми**

main.py:

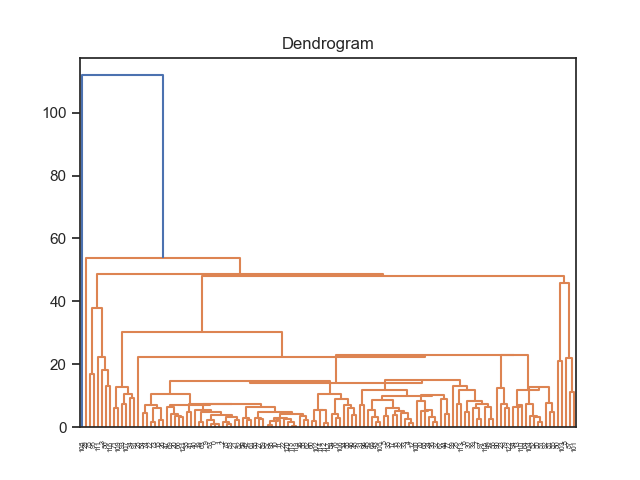
from scipy.spatial.distance import pdist  
from scipy.cluster.hierarchy import \*  
from matplotlib import pyplot as pl  
from statistics import mean  
import numpy as np  
import seaborn as sb  
import pandas as pd  
  
def Dispersion(Coords, centre, metric):  
 if len(Coords) < 2:  
 return 0  
 Distances = []  
 for i in range(len(Coords)):  
 Distances.append(list(pdist([Coords[i], centre], metric=metric))[0])  
 return sum([j \* j for j in Distances])/len(Distances)  
  
def GeometricDistancesToPoint(Coords, centre):  
 Distances = []  
 for i in range(len(Coords)):  
 vec = [centre[k] - Coords[i][k] for k in range(len(Coords[i]))]  
 Distances.append(pow(sum([j \* j for j in vec]), 0.5))  
 return Distances  
  
def GeometricDistancesSelf(Coords):  
 Distances = pdist(Coords, metric="euclidean")  
 return Distances  
  
Metrics = ["seuclidean", "cityblock", "chebyshev"]  
Methods = ["single", "complete", "centroid"]  
  
datasetFull = pd.read\_excel("./datasetFull.xlsx")  
dataset = datasetFull.drop(["Col1"], axis=1)  
sb.set\_theme(style="ticks")  
sb.pairplot(dataset)  
pl.savefig("dataset.png")  
pl.show()  
  
temp\_data = dataset.drop(["Code"], axis=1)  
coeffs = [[0 for i in Metrics] for j in Methods]  
for i in range(len(Metrics)):  
 distances = pdist(temp\_data, Metrics[i])  
 for j in range(len(Methods)):  
 linkageM = linkage(distances, method=Methods[j], metric=Metrics[i])  
 cophCoeff, cophMatrix = cophenet(linkageM, distances)  
 coeffs[i][j] = cophCoeff  
  
coeffsNP = np.array(coeffs)  
print("Таблиця кофенетичних кореляційних коефіцієнтів:")  
print(pd.DataFrame(coeffsNP, index=Metrics).set\_axis(Methods, axis=1), "\n")  
IbestMetric, IbestMethod = np.where(coeffsNP == coeffsNP.max())  
print("Найкращий з них:", round(coeffsNP.max(), 6), "з індексами", [IbestMetric[0], IbestMethod[0]])  
print("Обрана метрика:", Metrics[IbestMetric[0]])  
print("Обраний метод:", Methods[IbestMethod[0]], "\n")  
  
distances = pdist(temp\_data, Metrics[IbestMetric[0]])  
linkageM = linkage(distances, method=Methods[IbestMethod[0]], metric=Metrics[IbestMetric[0]])  
  
dendrogram(linkageM)  
pl.title("Dendrogram")  
pl.savefig("dendrogram.png")  
pl.show()  
  
clusterAmounts = [[], []]  
correctDist = 0  
for i in range(round(min(distances)), round(mean(distances)) + 1, 1):  
 clusters = fcluster(linkageM, i, criterion="distance")  
 clusterAmounts[0].append(i)  
 clusterAmounts[1].append(max(clusters))  
  
pl.plot(clusterAmounts[0], clusterAmounts[1], marker='o')  
pl.xlim(-1, round(max(clusterAmounts[0])) + 3)  
pl.ylim(0, round(max(clusterAmounts[1])) + 10)  
pl.xlabel("Distance")  
pl.ylabel("Amount of Clusters")  
pl.grid()  
pl.show()  
  
clusters = fcluster(linkageM, 10, criterion="maxclust")  
  
print(clusters, max(clusters))  
  
dataset["Clusters"] = clusters  
datasetFull["Clusters"] = clusters  
dataset = dataset.sort\_values(['Clusters'], ascending=[True])  
datasetFull = datasetFull.sort\_values(['Clusters'], ascending=[True])  
  
print("\nКількість елементів у кластерах:")  
print(dataset.groupby(['Clusters'])['Clusters'].count())  
  
print("\nКоординати центрів кластерів:")  
Mean1 = dataset.groupby(['Clusters'])['Index1'].mean()  
Mean2 = dataset.groupby(['Clusters'])['Index2'].mean()  
Mean3 = dataset.groupby(['Clusters'])['Index3'].mean()  
Centers = pd.DataFrame([Mean1, Mean2, Mean3]).transpose()  
print(Centers)  
  
Centers['Clusters'] = ["C" + str(i) for i in range(1, len(Centers) + 1)]  
dataset = pd.concat([dataset, Centers], ignore\_index=True, axis=0)  
sb.pairplot(dataset,  
 hue="Clusters",  
 palette="tab10",  
 markers=[\*["o" for i in range(len(Centers))], \*["D" for i in range(len(Centers))]])  
pl.savefig("datasetColored.png")  
pl.show()  
  
disps = []  
distsToCentres = []  
for i in range(1, len(Centers)+1):  
 tempTable = pd.DataFrame(dataset.drop(["Code"], axis=1)[dataset['Clusters'] == i]).drop(["Clusters"], axis=1)  
 tempCentres = Centers.drop(['Clusters'], axis=1)  
  
 disp = Dispersion(tempTable.values.tolist(), tempCentres.values.tolist()[i-1], Metrics[IbestMetric[0]])  
 disps.append(disp)  
  
 dTC = GeometricDistancesToPoint(tempTable.values.tolist(), tempCentres.values.tolist()[i-1])  
 distsToCentres.append(dTC)  
distsBeetwCentres = GeometricDistancesSelf(tempCentres)  
  
print("\nВнутрішньокластерна дисперсія:")  
pdDisps = pd.DataFrame(disps)  
pdDisps.set\_axis([i for i in range(1, len(Centers)+1)], inplace=True)  
pdDisps.index.names=['Clusters']  
pdDisps.rename(columns={0: "Within-cluster Dispersion"}, inplace=True)  
print(pdDisps)  
  
print("\nГеометричні відстані від елементів до центрів кластерів:")  
pdDTC = pd.DataFrame(distsToCentres)  
pdDTC.set\_axis([i for i in range(1, len(Centers)+1)], inplace=True)  
pdDTC.index.names=['Clusters']  
print(pdDTC)  
  
print("\nГеометричні відстані між центрами кластерів:")  
pdDBC = pd.DataFrame(distsBeetwCentres)  
pdDBC.rename(columns={0: "Distance"}, inplace=True)  
print(pdDBC)  
  
datasetFull.to\_excel("result.xlsx")

**5. Результати роботи програмного забезпечення (частина)**

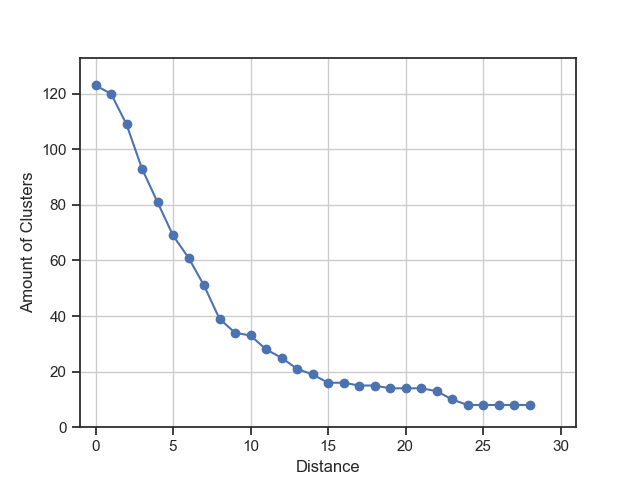
Візуалізація даних:



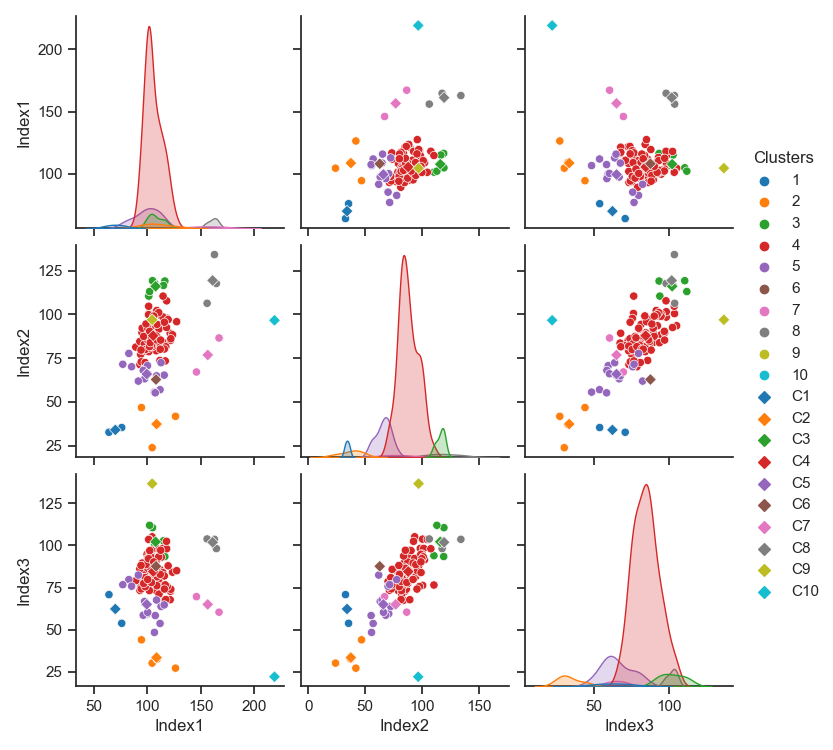
Дендрограма обраного методу:



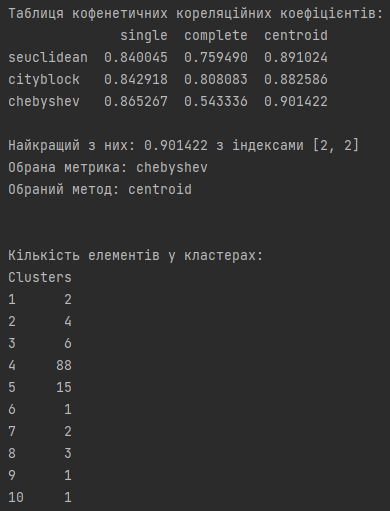
Графік для аналізу методом «ліктя»:

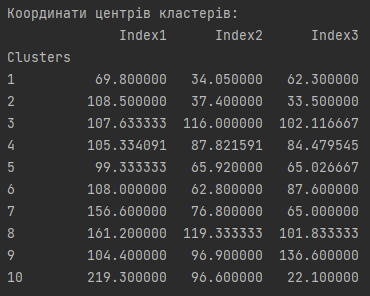


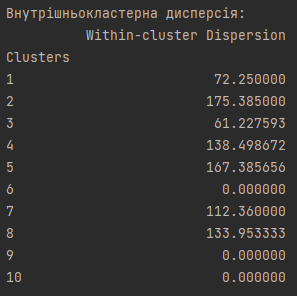
Візуалізація кластерів та їх центрів:

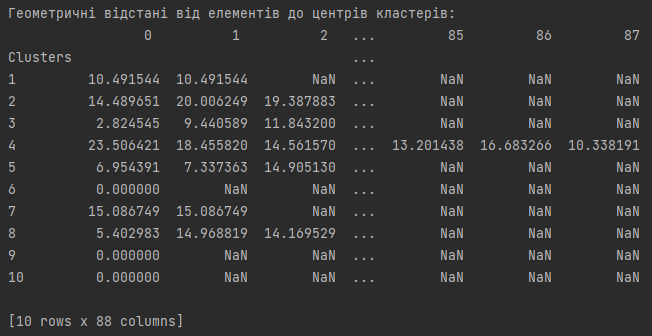


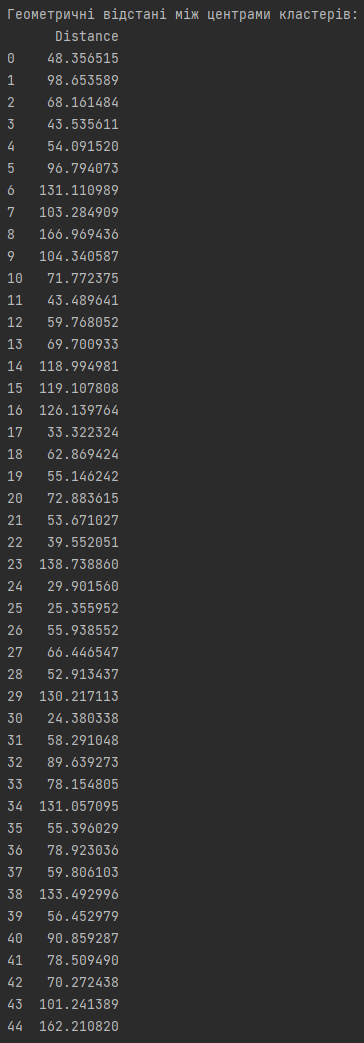
Вивід у консоль:











**6.1. Контрольні питання**

1. В чому полягає задача кластерного аналізу?

*Кластерний аналіз призначений для розбиття сукупності об'єктів на однорідні групи (кластери чи класи).*

1. Для яких задач обробки експериментальних даних використовуються методи ієрархічного кластерного аналізу?

*В галузі медицини – кластеризація захворювань, лікування захворювань або симптомів захворювань призводить до таксономій, що широко використовуються. В області психіатрії правильна діагностика кластерів симптомів, таких як параноя, шизофренія тощо, є вирішальною для успішної терапії. В археології за допомогою кластерного аналізу дослідники намагаються встановити таксономію кам'яних знарядь, похоронних об'єктів тощо. Відомі широкі застосування кластерного аналізу у маркетингових дослідженнях. Загалом, щоразу, коли необхідно класифікувати "гори" інформації до придатним подальшої обробки групам, кластерний аналіз виявляється дуже корисним і ефективним.*

1. Наведіть основні міри порівняння об'єктів між собою.

* *Евклідова.*
* *Стандартизована Евклідова.*
* *Міста.*
* *Махаланобіса.*
* *Мінковського.*
* *Чебишева.*

1. Що таке дендрограма?

*Дендрограма ілюструє, як складається кожен кластер шляхом малювання U-подібного зв’язку між неоднокомпонентним кластером та його дочірніми. Верхня частина U-посилання вказує на злиття кластерів. Дві ніжки U-подібного зв’язку вказують, які кластери були об’єднані. Довжина двох ніжок U-подібної ланки представляє відстань між дочірніми кластерами. Це також кофенетична відстань між вихідними спостереженнями в двох дочірніх кластерах.*

1. Що являють собою ієрархічні агломеративні методи кластерного аналізу?

*Ця група методів характеризується наступним об'єднанням вихідних елементів і відповідним зменшенням числа кластерів.*

*На початку роботи алгоритма всі об'єкти є роздільними кластерами. На першому етапі найбільш схожі об'єкти об'єднуються в кластер. На останніх кроках об'єднання продовжується до тех пор, поки всі об'єкти не будуть складати один кластер.*

1. Що являють собою ієрархічні дивізимні методи кластерного аналізу?

*Ці методи є логічною протилежністю агломеративним методам. На початку роботи алгоритму всі об'єкти належать одному кластеру, який на наступних кроках ділиться на менші кластери, у результаті утворюється послідовність груп, що розщеплюють.*

1. Наведіть основні способи зв'язування об'єктів у кластери.

* *Ближчого сусіда.*
* *Дальшого сусіда.*
* *Середнього зв'язку.*
* *Центроїдний.*
* *Медіанного зв'язку.*

1. Що таке кофенетичний кореляційний коефіцієнт?

*Кофенетичний кореляційний коефіцієнт – міра того, наскільки точно дендрограма зберігає попарні відстані між вихідними немодельованими точками даних.*

1. В чому полягають основні етапи ієрархічного кластерного аналізу?

* *Етап 1. Обчислення матриці відстанєй між об'єктами D.*
* *Етап 2. Зв'язування або групування об'єктів в бінарні ієрархічні дерева (дендрограми).*
* *Етап 3. Оцінка якості кластеризації.*
* *Етап 4. Виділення значущих кластерів.*
* *Етап 5. Візуалізація і аналіз значущих кластерів.*

1. Яким чином визначити значущу кількість кластерів?

*Найпопулярніший метод – метод «ліктя». Розраховується дистанція для кожної кількості кластерів та відображається у вигляді графіка, а користувач шукає зміну нахилу від крутого до пологого, візуальне «коліно», щоб визначити оптимальну кількість кластерів.*

1. З якою метою використовують функції MATLAB: pdist, square-form, linkage, dendrogram, cophenet, cluster, gscatter?

*Основні етапи ієрархічного кластерного аналізу в MATLAB*

* *pdist, squareform – Обчислення матриці відстанєй між об'єктами D.*
* *linkage, dendrogram – Зв'язування або групування об'єктів в бінарні ієрархічні дерева (дендрограми).*
* *сophenet – Оцінка якості кластеризації.*
* *cluster – Виділення значущих кластерів.*
* *gscatter – Візуалізація і аналіз значущих кластерів.*

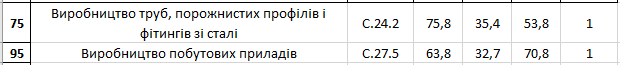
**6.2. Висновки**

В результаті виконання лабораторної роботи ми познайомилися з ієрархічним кластерним аналізом даних.

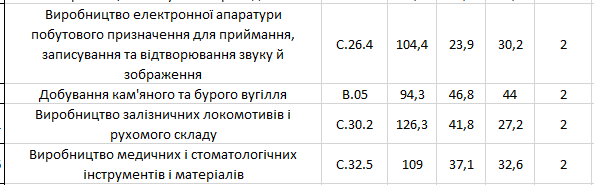
У ході аналізу представленого набору даних було виявлено, що цей набір даних доволі однорідний і навіть при оптимальній кількості кластерів залишився один великий кластер, у якому знаходиться майже 70% усіх записів. Але, після застосування методу «ліктя» було розглянуто сусідні значення від оптимального, таким чином було знайдено значення у 10 кластерів, яке при невеликій кількості нових малих кластерів дало змогу трошки поділити основний великий кластер на два менших.

Коротко про кластери:

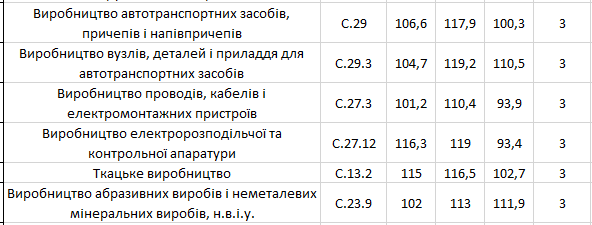
1. Перший кластер – загалом маленькі значення:



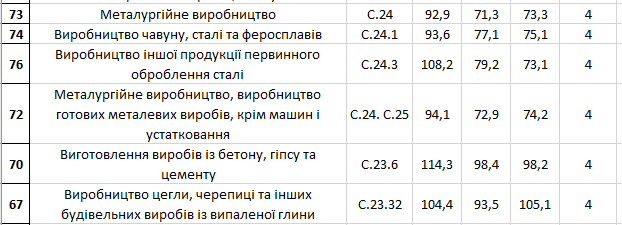
1. Другий кластер – маленькі значення двох параметрів:



1. Третій кластер – набір великих значень, порівняно із середніми:



1. Четвертий кластер – найбільший, середні значення. Що цікаво та цілком логічно – у цей кластер потрапили загальні значення, по узагальненим типам промисловостей:



…

1. П’ятий кластер – значення нижче середнього, той кластер, що відділився від 4го:



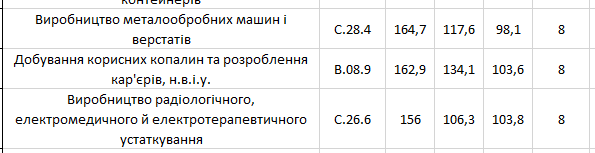
1. Шостий кластер – один з двох «кластерів-аномалій», що містить тільки один елемент:



1. Сьомий кластер – дані з високими значеннями приросту Червень до Травня 2015 і низкими іншими:



1. Восьмий кластер – дані з високими значеннями приросту Червень до Травня 2015 і високими/середніми іншими:



1. Дев’ятий кластер – один з двох «кластерів-аномалій», що містить тільки один елемент:



1. Десятий кластер – дані, що мають дуже високий приріст Червень до Травня 2015:

