Навчально-методична література

Міністерство освіти і науки України
Тернопільський національний технічний університет ім. Івана Пулюя
Кафедра комп'ютерних наук

Методичні вказівки
до лабораторної роботи №5 з курсу
«Інтелектуальний аналіз даних»
для студентів спеціальності 122 «Комп'ютерні науки»

Тема: «Побудова моделі кластеризації»

Тернопіль – 2017

Побудова моделі кластеризації. Методичні вказівки до лабораторної роботи з навчальної дисципліни «Інтелектуальний аналіз даних» для студентів спеціальності 122 "Комп'ютерні науки", кафедра КН ТНТУ ім. І.Пулюя, Тернопіль, 2017 р.

Укладач:

ст. викл. кафедри КН ТНТУ ім. І. Пулюя ***Козбур Галина Володимирівна***

Відповідальний за випуск:

ст. викл. кафедри КН ТНТУ ім. І. Пулюя ***Козбур Галина Володимирівна***

Лабораторна робота № 5

Тема: Побудова моделі кластеризації

Мета роботи: освоєння методу побудови моделі кластеризації.

Завдання: навчитись виконувати кластеризацію даних для виявлення неочевидних закономірностей та відмінностей у них, використовуючи надбудови інтелектуального аналізу даних MS SQL Server (для MS Office Excel або StatSoft STATISTICA).

Порядок виконання роботи

1. Для виконання даної лабораторної роботи потрібно попередньо встановити Microsoft SQL Server (версія >=2008). Файли інсталяції завантажити з офіційного сервера Microsoft; можна використати посилання:

- https://download.microsoft.com/download/4/C/7/4C7D40B9-BCF8-4F8A-9E76-06E9B92FE5AE/ENU/x64/SQLFULL_x64_ENU_Install.exe

- https://download.microsoft.com/download/4/C/7/4C7D40B9-BCF8-4F8A-9E76-06E9B92FE5AE/ENU/x64/SQLFULL_x64_ENU_Lang.box

- https://download.microsoft.com/download/4/C/7/4C7D40B9-BCF8-4F8A-9E76-06E9B92FE5AE/ENU/x64/SQLFULL_x64_ENU_Core.box

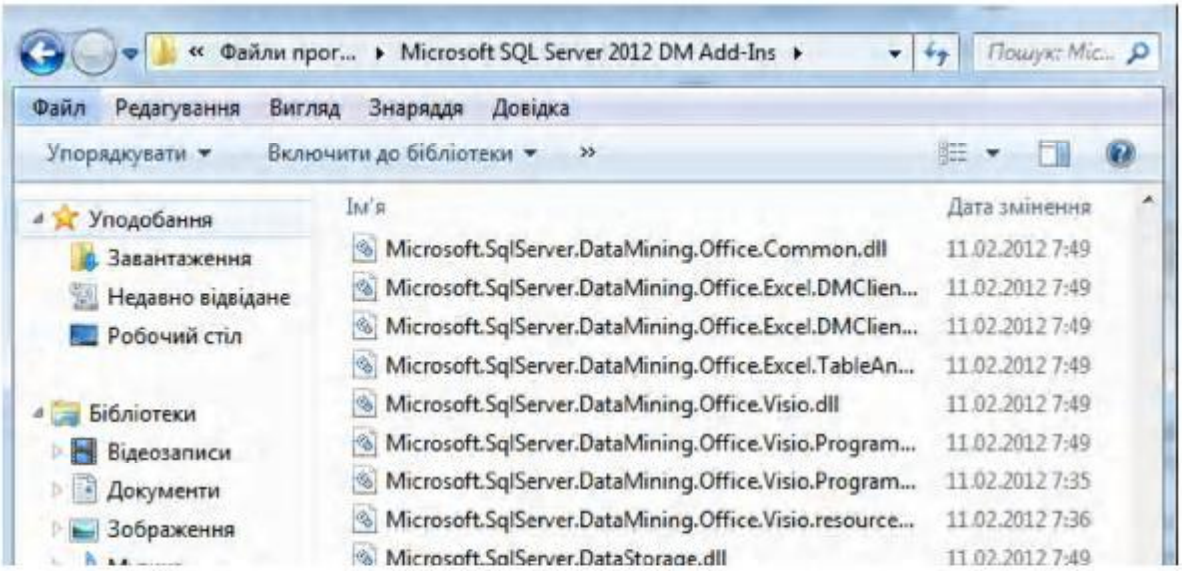
При інсталяції Microsoft SQL Server вибрати версію Free trial evaluation.

2. Безпосередньо лабораторну роботу можна виконати:

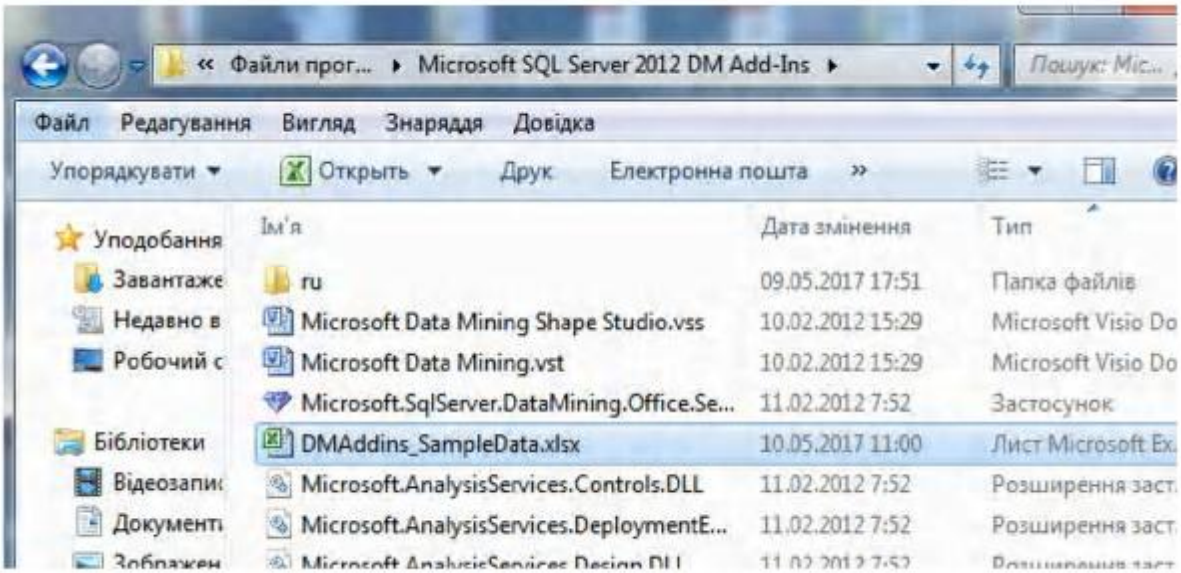
- в надбудові MS Office Excel (Клієнт інтелектуального аналізу даних для Excel – надбудови інтелектуального аналізу даних SQL Server: https://download.microsoft.com/download/5/6/4/5649CD1C-FB20-4A1C-B5FA-4ABBF8CE23FB/RUS/x64/SQL_AS_DMAddin.msi (64-біт), **Ошибка! Недопустимый объект гиперссылки.** (32-біт)) або

- в прикладному пакеті StatSoft STATISTICA (система для статистичного аналізу даних – https://support.quest.com/statistica/13.2/download-new-releases).

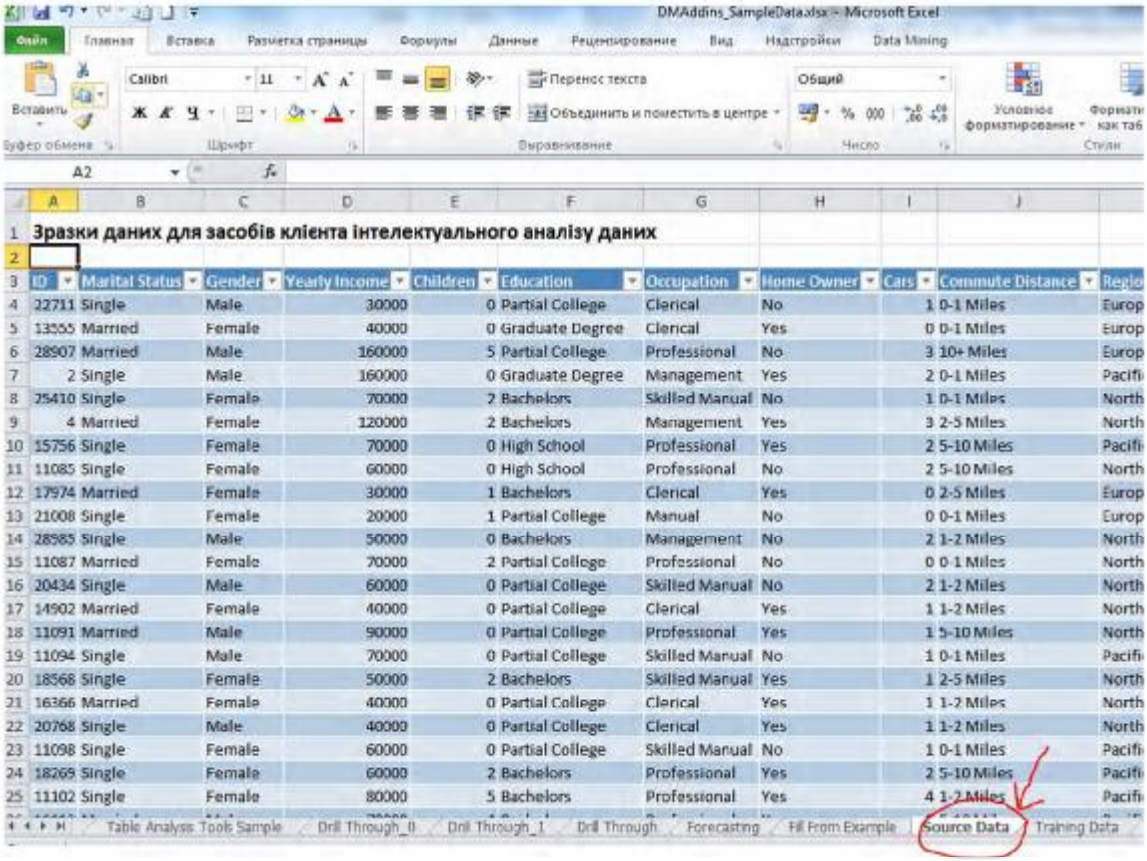
Обидва програмні засоби вимагають середовища Microsoft SQL Server. Окрім цього, надбудова Excel буде вимагати установки MS Visio відповідної версії, сумісної зі встановленою версією MS Office (див. скріншот).



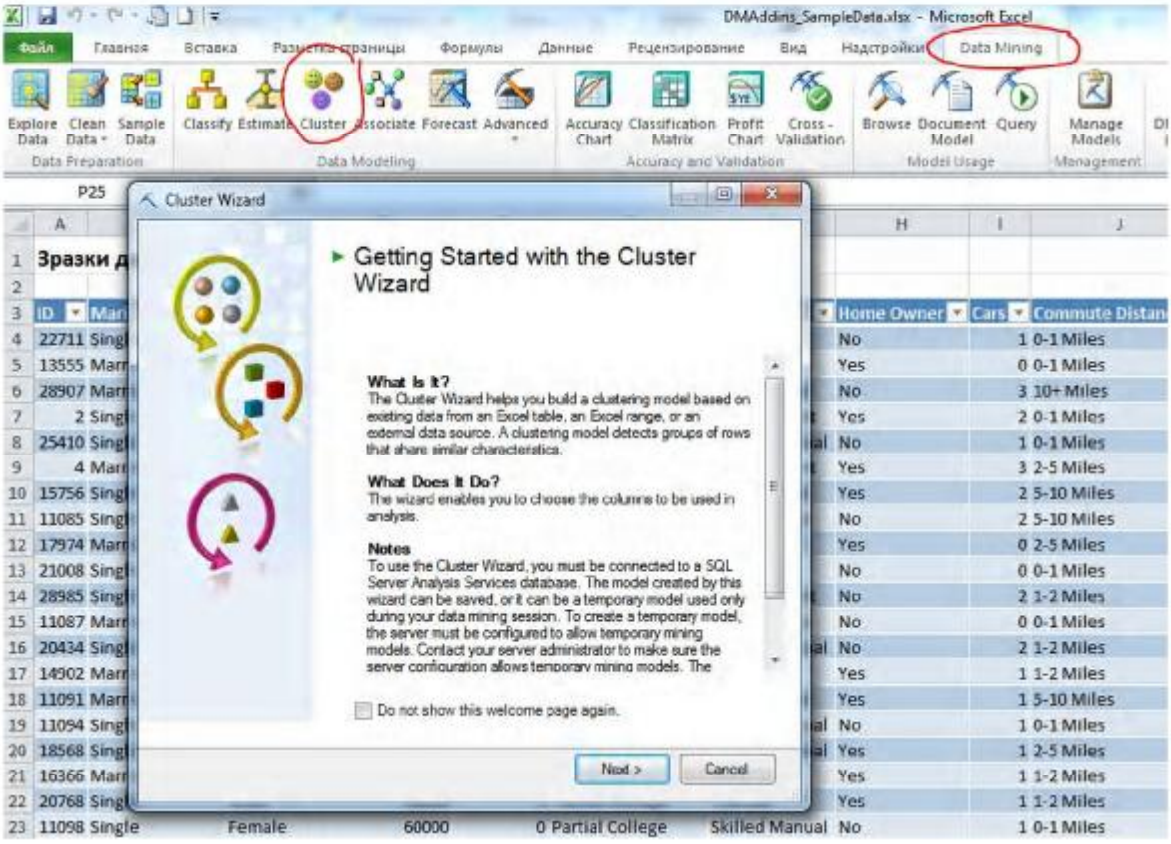
При інсталяції надбудови SQL Data Mining для Excel будуть встановлені файли із тренувальними наборами даних (див. скріншот).



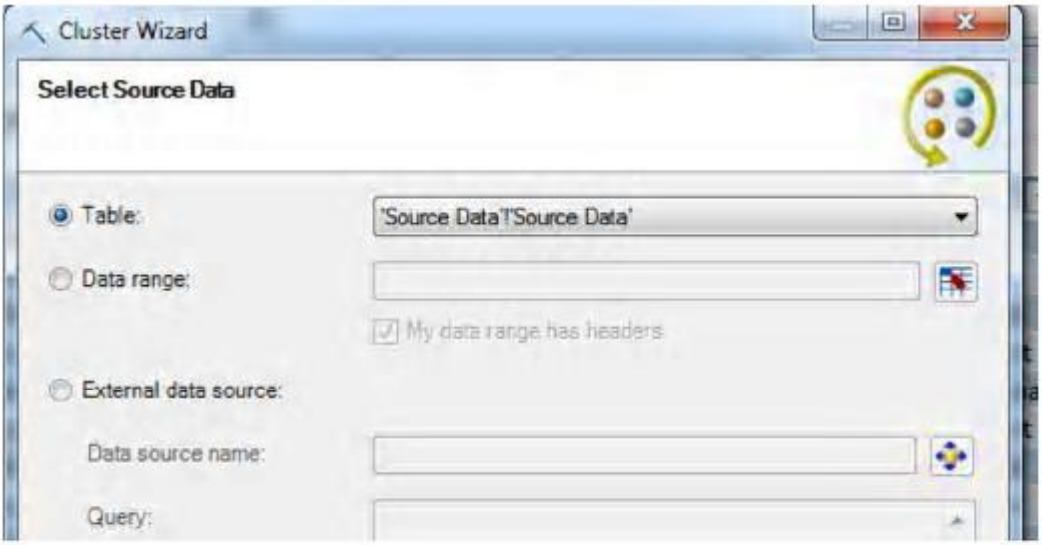
3. Розглянемо основні етапи побудови моделі кластеризації даних. Відкриваємо відзначений у попередньому скріншоті файл **DMAddins_SampleData.xlsx**, вибираємо лист з тренувальним набором даних, наприклад, **Source Data** (див скріншот).



4. У розділі Data Mining вибираємо майстер кластеризації Cluster (скріншот).

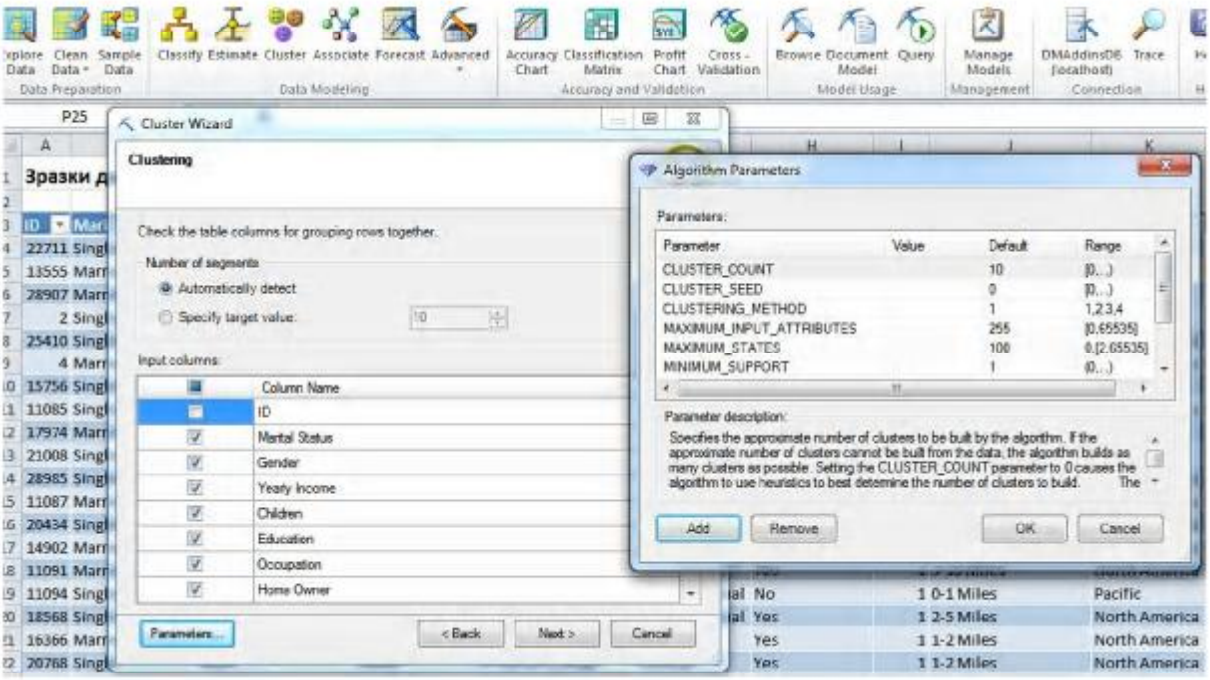


Вказуємо джерело даних для кластеризації:

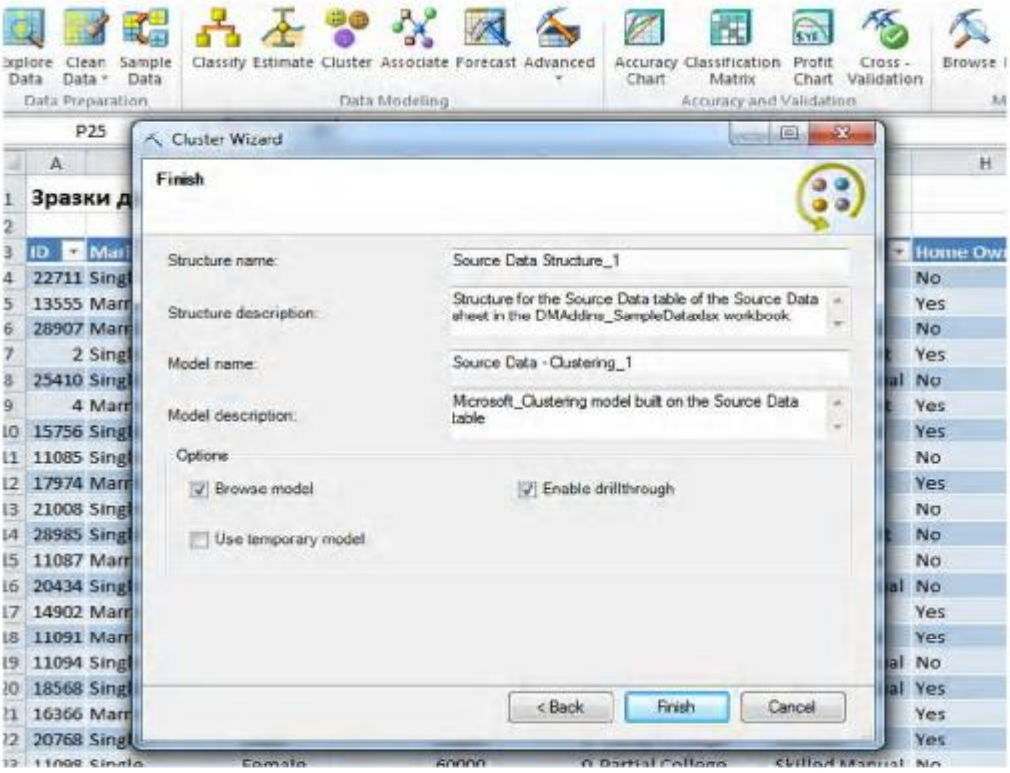


Виключаємо з розгляду атрибути, які не мають смислового значення для класифікації об'єктів (в даному випадку – ID записів) та ознайомлюємось із параметрами побудови моделі кластеризації (скріншот).

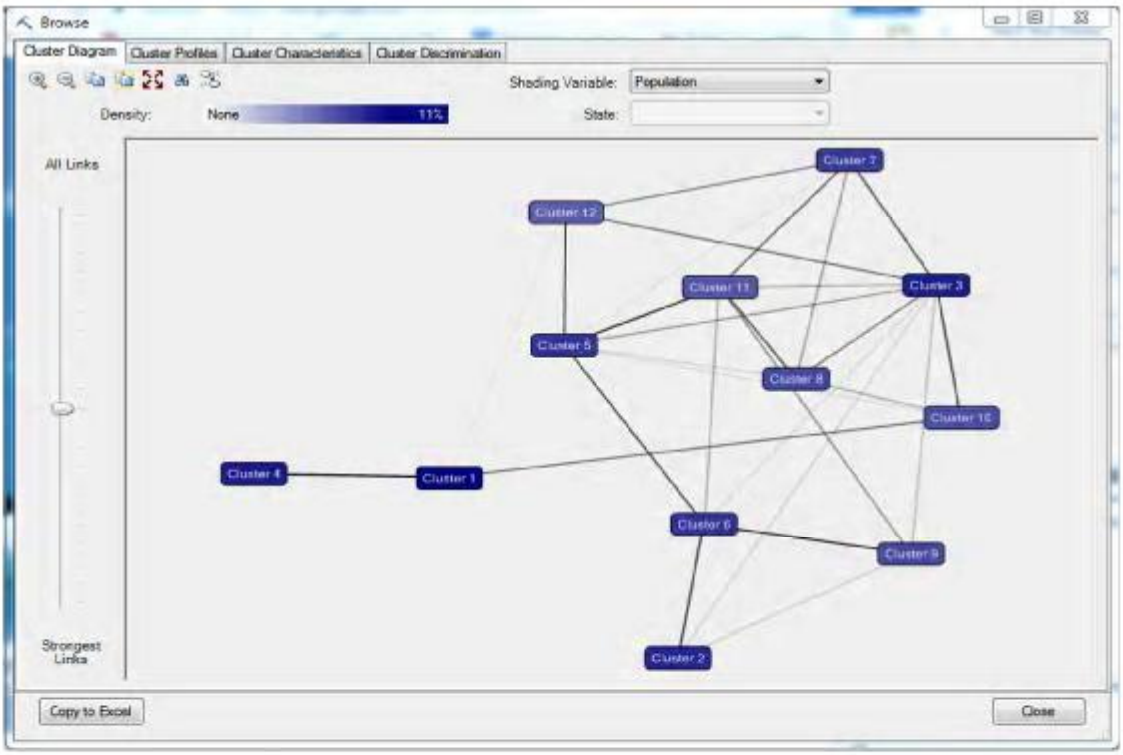
Кількість кластерів для розбиття об'єктів залишаємо без змін (автоматичний вибір).



Залишаємо без змін відсоток даних (чи максимальну кількість рядків таблиці), відібраних для тестування моделі, та завершуємо стапи побудови кластерів для тренувальної моделі (скріншот). У таблиці «Finish» можемо змінити заголовки і коментарі.

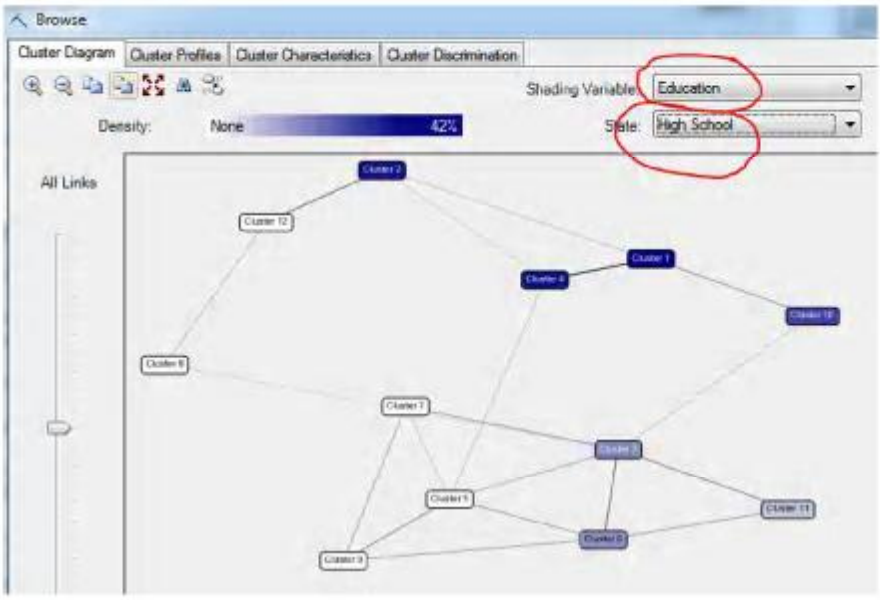


5. Після автоматичного створення моделі кластерного аналізу отримуємо вікно із побудованою діаграмою. Це означає, що кластеризацію тренувального набору даних проведено. Наступний крок – аналіз результатів.



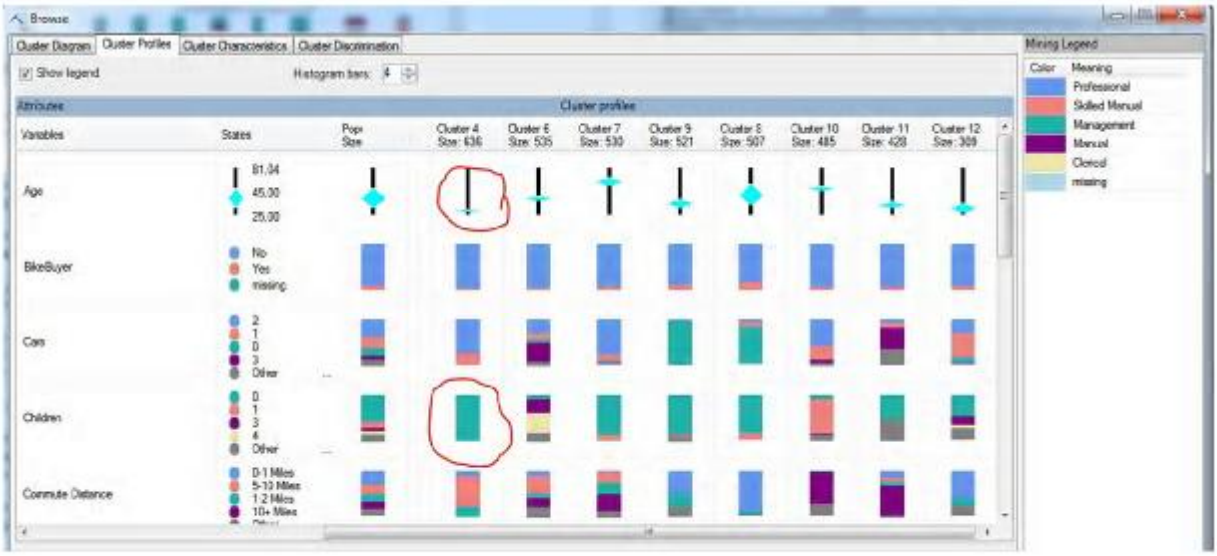
Як видно з діаграми, алгоритм автоматично знайшов у наборі даних 12 різних «угруповань», або кластерів. Чим інтенсивніше виділена лінія, що пов'язує два кластери, тим більшою є схожість цих кластерів. Щоб виділити найсильніші з цих зв'язків, потрібно перетягнути повзунок, що ліворуч від діаграми.

Інтенсивність заливки самого кластера відображає рівень підтримки – чим більше об'єктів увійшло до кластера, тим інтенсивніша заливка. Підвівши курсор до кластера, у виринаючому вікні побачимо коментар. Щоб побачити, як розподілились за кластерами об'єкти різних категорій (за віком, статтю тощо), потрібно змінити значення атрибутів та їх стану:



6. Вкладка «Профілі кластерів» дозволяє порівняти склад усіх кластерів. Порівнюючи наповнення кластерів (для зручності порівняння клікайте на назві кластера, програма буде сортувати змінні кластерів від найбільш відмінних до найбільш подібних чи навпаки), виявляємо їх відмінні риси, або особливості, та перейменовуємо. Наприклад, до кластеру 4 увійшли наймолодші за віком люди, які не мають дітей. Тому кластер 4 можна назвати «Бездітна молодь».

Для перейменування кластера викликаємо контекстне меню на полі з номером та розміром кластера.

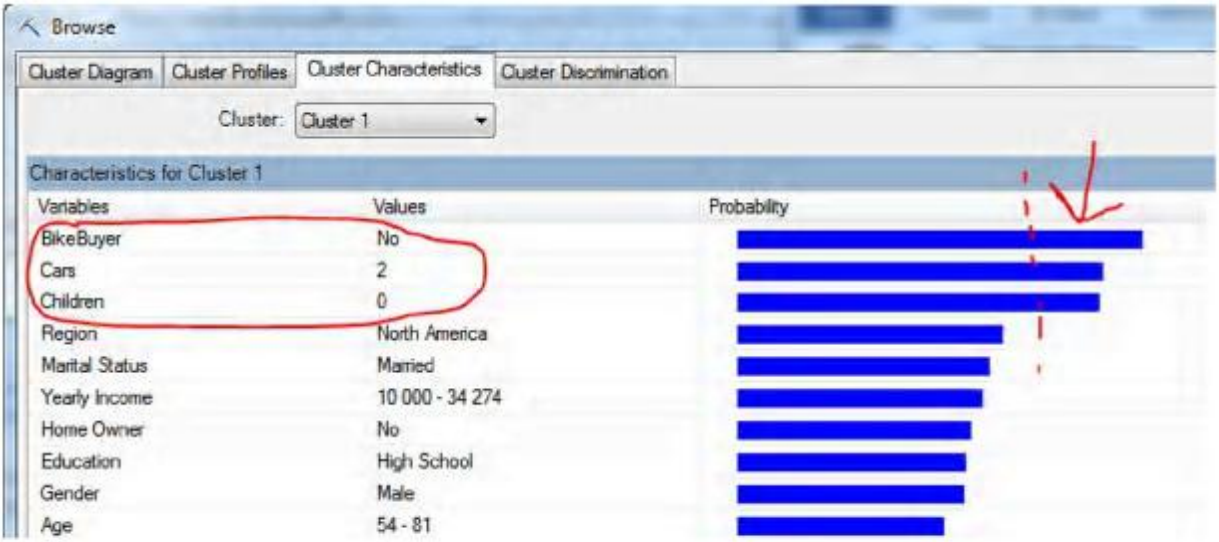


Даємо назви усім кластерам. Так з'являться, наприклад, категорії «Неодружені з дипломом бакалавра» або «Сімейні середнього віку, власники 2-3 машин» тощо. Для зручності можна тимчасово приховувати деякі стовпці.

Ознайомлюємось із додатковими можливостями вкладки «Профілі кластерів».

7. У вкладці «Характеристики кластерів» ознайомлюємось детальніше із якісним та кількісним наповненням всієї генеральної сукупності (популяції) або кластерів окремо. Виділяємо найзначущіші характеристики кластерів (за найбільшим відсотковим вмістом). Зауважимо, що ці характеристики повинні узгоджуватись із назвами, наданими кластерам на попередньому кроці.

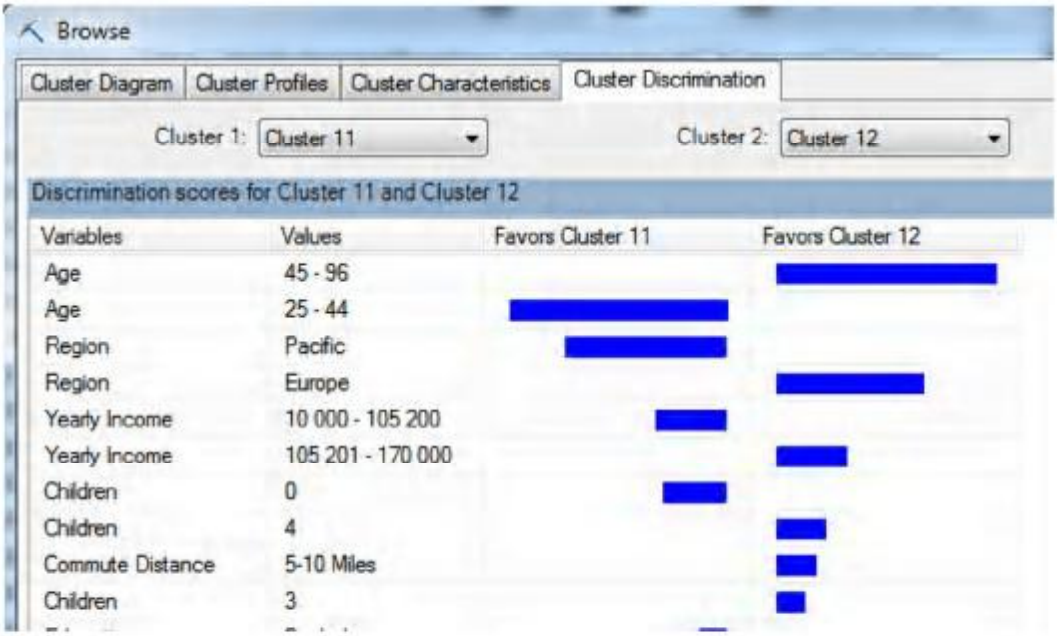
Так, наприклад, за даними для кластеру 1, показаними на скріншоті, можна очікувати назву кластеру як «Бездітні власники 2-х машин, що не купують велосипедів».



8. За допомогою вкладки «Порівняння кластерів» можемо порівнювати атрибути між кластером та всією популяцією чи між двома будь-якими кластерами. Перебором пар кластерів вибираємо пару (або 2-3 пари), в яких відмінні риси проявлені найяскравіше.

Наприклад, наступний скріншот показує, що алгоритм відокремив у кластери 11 і 12 людей за віком (11-й кластер: середній вік і молодші – це, наприклад, люди, що зростають у своїй професійній майстерності, тоді як 12-й кластер: люди старшого віку – це, наприклад, люди, що досягли піку своєї професійної

майстерності та пенсіонери) та, аналогічно, за місцем проживання (європейці та жителі островів чи узбережжя Тихого океану). Менші за важливістю фактори кластеризації – річний дохід та кількість дітей у сім'ї.



Повернемось до вкладки «Характеристики кластерів», для кожного з цих кластерів зробимо копію в Excel (на одному листі). Для цього використовуємо спеціальну кнопку «Copy to Excel». Отримуємо, наприклад:

	A	B	C	D	E	F	G
1	Source Data - Clustering_6				Source Data - Clustering_6		
2	Cluster Characteristics				Cluster Characteristics		
3	Cluster 11				Cluster 12		
4	Variables	Values	Probability		Variables	Values	Probability
5	BikeBuyer	No	85 %		BikeBuyer	No	93 %
6	Region	Pacific	84 %		Yearly Income	77 748 - 152 691	91 %
7	Home Owner	Yes	68 %		Region	Europe	77 %
8	Yearly Income	77 748 - 152 691	66 %		Age	54 - 81	72 %
9	Occupation	Professional	65 %		Marital Status	Married	70 %
10	Education	Bachelors	61 %		Home Owner	Yes	70 %
11	Commute Distance	10+ Miles	60 %		Occupation	Management	59 %
12	Age	25 - 37	60 %		Cars	4	54 %
13	Marital Status	Married	59 %		Gender	Female	51 %
14	Gender	Female	52 %		Gender	Male	49 %
15	Children	5	52 %		Children	4	43 %
16	Gender	Male	48 %		Occupation	Professional	41 %
17	Children	0	46 %		Commute Distance	10+ Miles	37 %
18	Marital Status	Single	41 %		Cars	3	34 %
19	Cars	3	41 %		Commute Distance	5-10 Miles	31 %
20	Age	38 - 45	39 %		Home Owner	No	30 %
21	Cars	4	35 %		Marital Status	Single	30 %
22	Occupation	Management	35 %		Education	High School	29 %
23	Home Owner	No	32 %		Education	Partial College	26 %

Зауважимо, що в Excel будуть відображені базові дані, що використовувались на попередніх етапах майстром кластеризації. Повернувшись до електронних таблиць, можемо далі виконувати з отриманими результатами усі екселівські маніпуляції: сортувати, фільтрувати, застосовувати інші алгоритми аналізу даних для глибшого аналізу, тощо.

9. Переходимо безпосередньо до аналізу даних: знаходимо неочевидні зв'язки та закономірності між кластерами та всередині кластерів, робимо висновки.

Зауважимо, що обробку даних з допомогою майстра кластеризації необхідно робити «в один прохід»: при повторній обробці моделі алгоритм кластеризації запускається «з нуля», при цьому алгоритмом може бути вибрана інша кількість кластерів, хоча основні висновки для одного й того ж набору даних, отримані про різних проходах, очевидно, будуть узгоджуватись.

10. Робимо новий прохід майстра кластеризації для того ж навчального набору даних. При цьому змінюємо налаштування алгоритму: вибираємо бажану кількість кластерів (4-6), можливо, змінюємо кількість рядків бази даних для побудови моделі. Робимо аналіз новоствореної моделі кластеризації.

11. Порівнюємо результати двох побудованих моделей кластеризації – з автоматичними налаштуваннями алгоритму та бажаними, декларованими нами. Робимо висновки.

12. Підготовка власних даних (згідно з призначеним варіантом) та знаходження моделі кластеризації з розширеним аналізом результатів.

Зміст звіту по роботі:

1. Титульний лист.

2. Мета роботи.

3. Результати роботи з навчальними даними з поясненнями та скріншотами.

4. Побудова моделі кластеризації для **індивідуальної бази даних** для виявлення відмінностей, особливостей та інших неочевидних характеристик даних. Описати дані та зробити скріни з джерел(-а) інформації.

5. Аналіз результатів та формулювання висновків щодо можливого їх використання.

6. Висновки по роботі.

Індивідуальні завдання лабораторної роботи № 5:

Дані для кластерного аналізу беремо з ресурсу https://github.com/devua/csv .

Варіант 1. https://github.com/devua/csv/blob/master/job-search/job-search-2017.csv

Варіант 2. https://github.com/devua/csv/blob/master/salaries/2016_may_final.csv

Варіант 3. https://github.com/devua/csv/blob/master/portrait_2016.csv

Варіант 4. https://github.com/devua/csv/blob/master/salaries/2014_dec_final.csv

Варіант 5. https://github.com/devua/csv/blob/master/relocation/relocation_2016.csv

Варіант 6. https://github.com/devua/csv/blob/master/salaries/2016_dec_final.csv

Варіант 7. https://github.com/devua/csv/blob/master/results-of-the-year/results-of-the-2015.csv

Варіант 8. https://github.com/devua/csv/blob/master/salaries/2016_may_final.csv

Варіант 9. https://github.com/devua/csv/blob/master/results-of-the-year/results-of-the-2016.csv

Варіант 10. https://github.com/devua/csv/blob/master/salaries/2015_may_final.csv

Варіант 11. https://github.com/devua/csv/blob/master/salaries/2016_dec_final.csv

Варіант 12. https://github.com/devua/csv/blob/master/salaries/2014_may_final.csv

Варіант 13. https://github.com/devua/csv/blob/master/salaries/2015_may_final.csv

Варіант 14. https://github.com/devua/csv/blob/master/salaries/2015_dec_final.csv

Варіант 15. https://github.com/devua/csv/blob/master/salaries/2015_dec_final.csv