Лабораторна работа 2:

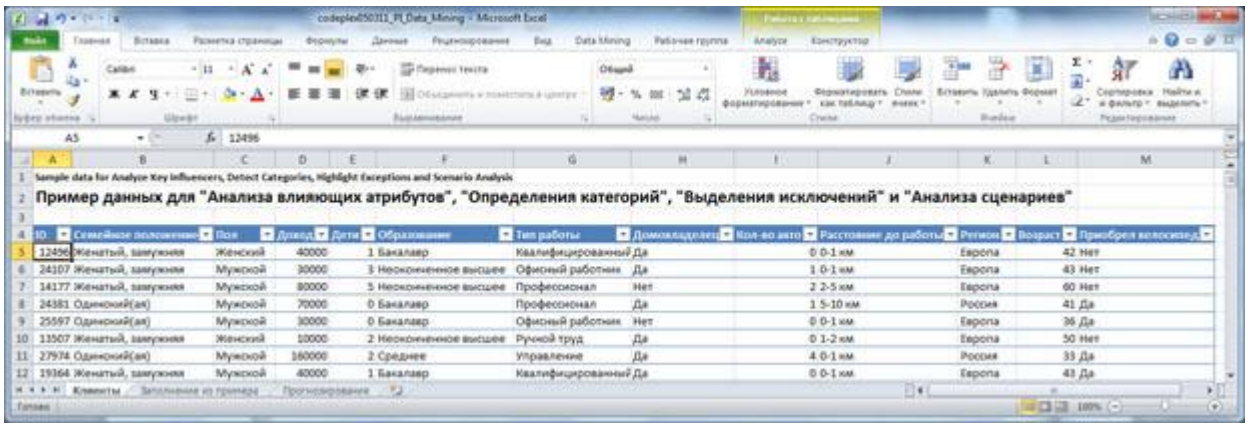
Використання інструментів "AnalyzeKeyInfluencers" і "DetectCategories"

Анотація: В ході даної лабораторної роботи буде розглянуто використання інструментів "Аналіз ключових факторів впливу" ("AnalyzeKeyInfluencers") і "Виявлення категорій" ("DetectCategories"), що відносяться до компоненту "Засоби аналізу таблиць для Excel" пакета надбудов інтелектуального аналізу даних для MicrosoftOffice 2007.

Ключові слова: таблиця, меню, SQL, server, заголовки, ПО, файл, інформація, ідентифікатор, параметр, предметної області, поле, значення, плата, поєднання, дискретизація, інтервал, діапазон, безлічі, атрибут, ID, визначення, відображення, діаграма

Почнемо безпосереднє вивчення інструментів інтелектуального аналізу даних (DataMining, сокр.DM). До складу пакета надбудов для MS Office 2007 входить електронна таблиця із зразками даних. Вона може бути відкрита з меню Пуск> Надбудови інтелектуального аналізу даних. Microsoft SQL Server 2008. Але переведено вміст файлу тільки частково - перша сторінка зі змістом і деякі заголовки. Тому в роботі буде використовуватися локалізований набір даних для аналізу, доступний для скачування за адресою http://russiandmaddins.codeplex.com/.

Скачайте файл, відкрийте його і відформатуйте дані на аркуші "клієнти" як таблицю (див. "Надбудови інтелектуального аналізу даних для MicrosoftOffice"). Перейдіть на вкладку Analyze (рис. 1). Анализируемая таблиця містить дані фірми, що продає велосипеди. У ній зібрано інформацію про клієнтів (ідентифікатор, сімейний стан, стать і т.д.) і зазначено, придбав клієнт велосипед чи ні.



Sample data for Analyze Key Influencers, Detect Categories, Highlight Exceptions and Scenario Analysis

Пример данных для "Анализа влияющих атрибутов", "Определения категорий", "Выделения исключений" и "Анализа сценариев"

ID	Семейное положение	Пол	Доход	Дети	Образование	Тип работы	Домовладелец	Кол-во авто	Расстояние до работы	Регион	Возраст	Приобрел велосипед
12496	Женатый, замужняя	Женский	40000	1	Бакалавр	Квалифицированный	Да	0	0-1 км	Европа	42	Нет
24107	Женатый, замужняя	Мужской	30000	3	Неоконченное высшее	Офисный работник	Да	1	0-1 км	Европа	43	Нет
14177	Женатый, замужняя	Мужской	80000	5	Неоконченное высшее	Профессионал	Нет	2	2-5 км	Европа	60	Нет
24381	Одинокий(ая)	Мужской	70000	0	Бакалавр	Профессионал	Да	1	5-10 км	Россия	41	Да
25597	Одинокий(ая)	Мужской	30000	0	Бакалавр	Офисный работник	Нет	0	0-1 км	Европа	36	Да
13507	Женатый, замужняя	Женский	10000	2	Неоконченное высшее	Ручной труд	Да	0	1-2 км	Европа	50	Нет
27974	Одинокий(ая)	Мужской	160000	2	Среднее	Управление	Да	4	0-1 км	Россия	33	Да
19364	Женатый, замужняя	Мужской	40000	1	Бакалавр	Квалифицированный	Да	0	0-1 км	Европа	43	Да

Рис. 1. Підготовлений набір даних

Аналіз ключових факторів впливу

Інструмент AnalyzeKeyInfluencers дозволяє визначити, як залежить цікавий для нас параметр від інших. При цьому важливо правильно визначити, що і від чого може залежати. Власне в цьому почасти й полягає майстерність аналітика, засноване на його знанні предметної області і використовуваних методів DM.

У зв'язку з тим, що ми оцінюємо ступінь взаємного впливу різних параметрів один на одного, варто відразу прибрати з розгляду повністю незалежні і навпаки, повністю залежні. Нехай, наприклад, ми хочемо оцінити вплив різних чинників на рівень заробітної плати людини. Якщо у нас є поле, що містить унікальний ідентифікатор (наприклад, порядковий номер запису в таблиці або номер паспорта), його варто прибрати з розгляду, що не впливає на значення досліджуваного параметра. Інший приклад, нехай у нас є значення заробітної плати за місяць і за рік, що розраховується як заробітна плата за місяць, помножена на 12. Ми знаємо, що ці значення завжди пов'язані, шукати залежність одного від іншого засобами DM не має сенсу, а наявна сильна залежність приховає вплив інших факторів, яке ми як раз і хочемо виявити.

Тепер визначимо, від чого залежить рішення клієнта про покупку велосипеда. Натискаємо на кнопку Analyze Key Influencers і вказуємо в якості цільового стовпчика стовпчик "Придбав велосипед" (рис. 2). Перейдемо по

посиланню "Choose columns to be used for analysis", щоб вказати параметри, вплив яких ми хочемо оцінити (рис. 3). Тут скинемо позначку навпроти "ID" та "Придбав велосипед" (хоча останнє можна і не робити).

Рис. 2. Вибір залежного параметра для аналізу

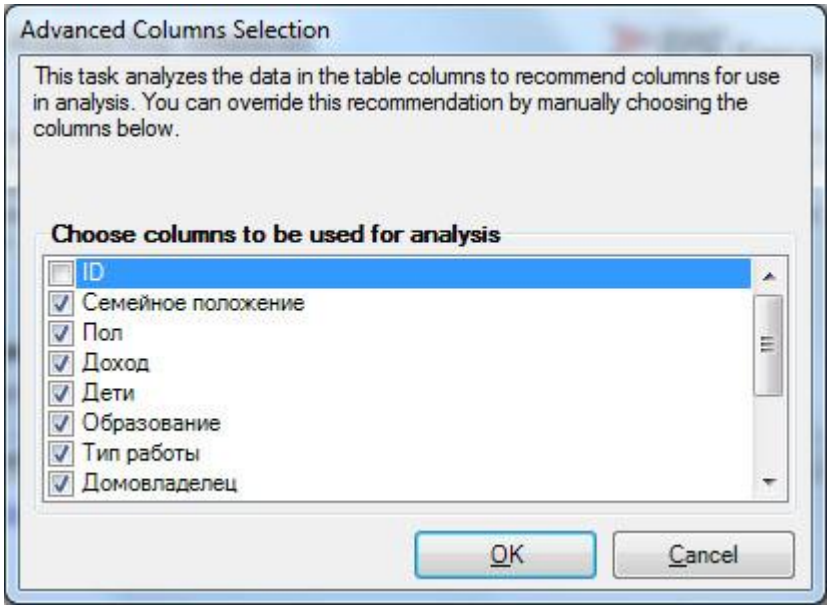


Рис. 3. Вибір параметрів, від яких залежить аналізований

Після запуску процедури аналізу (по кнопці Run, рис. 2) буде сформований звіт про фактори впливу та запропоновано формування додаткового порівняльного звіту (рис. 4). В основному звіті вказується стовпець (Column), його значення (Value), значення цільового стовпчика, з яким воно зв'язується (Favors) і рівень впливу (Relative Impact), оцінюваний пошкале від 0 до 100 балів. З представленногона рис. 4 звіту видно, що на рішення не купувати велосипед в найбільшою мірою впливає наявність 2-х автомобілів. У той же час не слід сприймати оцінку 100 балів, як ознака того, що в 100% випадків

власники 2-х машин велосипед не купували (подивіться набір даних, там є і поєднання "2 машини - велосипед куплений", але їх меншість). Другий за рівнем впливу на відмову від покупки фактор - "Сімейний стан" = "одружений, заміжня".

Найбільший вплив на позитивне рішення про придбання велосипеда надає відсутність у клієнта машини.

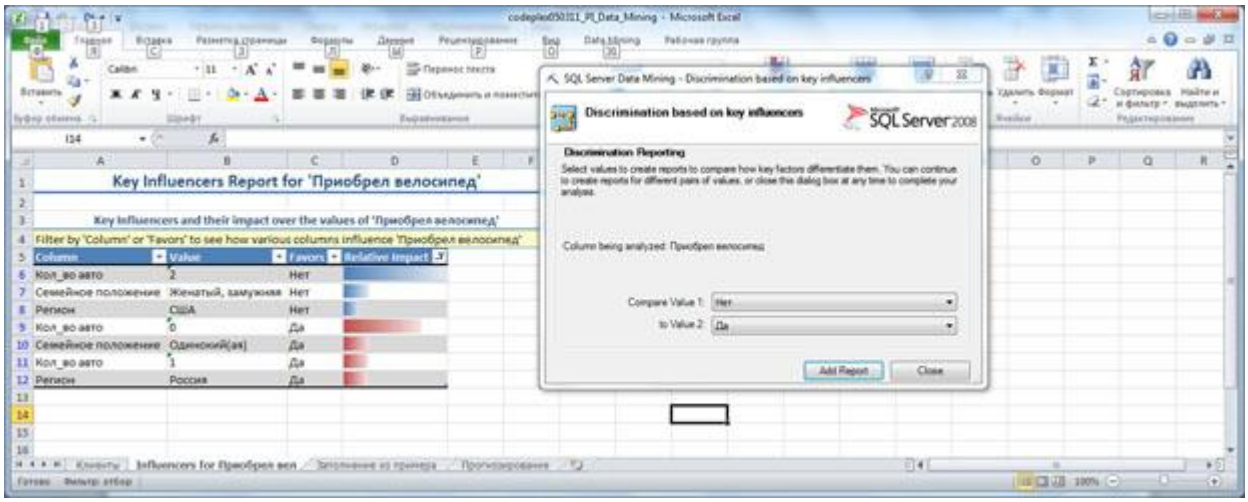


Рис. 4. Основний звіт

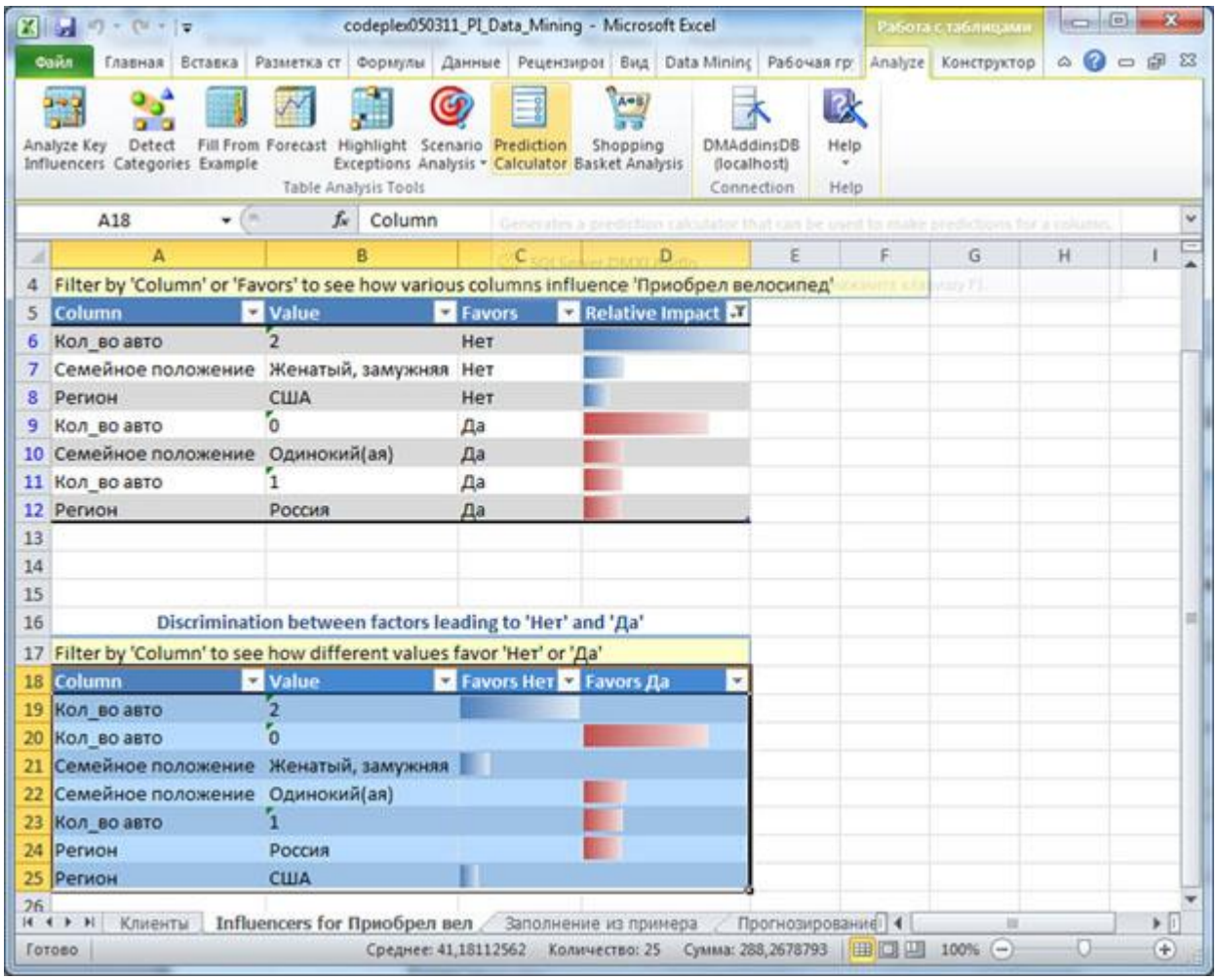


Рис.5. порівняльний звіт

Якщо додати порівняльний звіт для двох обраних значень (Рис.4, Add Report), можна побачити, чим відрізняється вибір на користь одного значення цільового стовпчика від вибору на користь іншого (Рис.5). У нашому прикладі просто відбудеться перегрупування вихідного звіту, тому що можливих значень всього 2. В інших випадках, додатковий звіт дозволяє провести детальне порівняння двох обраних варіантів.

Як зазначається в [1], якщо цільової або інший стовпець, що обробляється інструментом Analyze Key Influencers, містить багато різних числових значень, то проводиться дискретизація. Весь інтервал значень ділиться на кілька діапазонів, кожен з яких розглядається як одне з можливих значень (наприклад, замість точного значення 2,5 ми отримаємо "діапазон від 2 до 3").

Завдання 1. Проведіть аналіз відповідно до розглянутим прикладом.

Завдання 2. На тому ж наборі даних проаналізуйте залежність рівня доходу від освіти, сімейного стану, типу роботи, статі, віку та регіону проживання клієнта. Опишіть результати.

Доповніть звіт порівняльним аналізом для найнижчого і наступного за ним діапазону доходу. А потім - для найнижчого і найвищого діапазону. Опишіть результати проведеного аналізу та запропонуйте їх інтерпретацію.

Завдання 3. Запропонуйте свій варіант аналізу даних, і приклад використання отриманих результатів.

Сформований звіт буде доступний і в разі, якщо ви відкриєте файл і на іншому комп'ютері (без підключення каналітіческім службам SQLServer).

Щоб повернути дані в початковий стан потрібно видалити листи з сформованими звітами.

виявлення категорій

Інструмент Detect Categories дозволяє вирішити задачу кластеризації, тобто поділу всього безлічі варіантів на "природні" групи, члени яких найбільш близькі за рядом ознак. Подібна задача також називається завданням сегментації.

Отже, в нашому прикладі є опис безлічі клієнтів і потрібно розділити їх на невелику кількість груп (щоб окремим групам сформувати спеціальну пропозицію і т.п.).

У зв'язку з тим, що в процесі роботи інструмент додає дані в вихідну таблицю, рекомендується перед початком роботи зробити її копію (Рис.6).

Після цього натискаємо кнопку Detect Categories і налаштовуємо параметри (Рис.7). Тут хочеться звернути увагу на атрібутID, який як було зазначено вище, не можна буде враховувати в ході аналіза.Поетому він автоматично виключений. У нашому випадку, інші атрибути можна залишити. Ще раз хотілося б повторити, що цей вибір кожен раз робиться виходячи з особливостей предметної області.

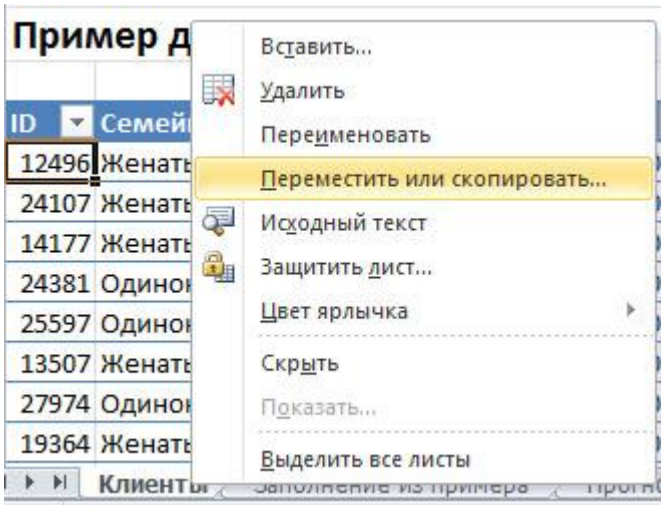


Рис.6. Перед початком роботи краще скопіювати лист Excel

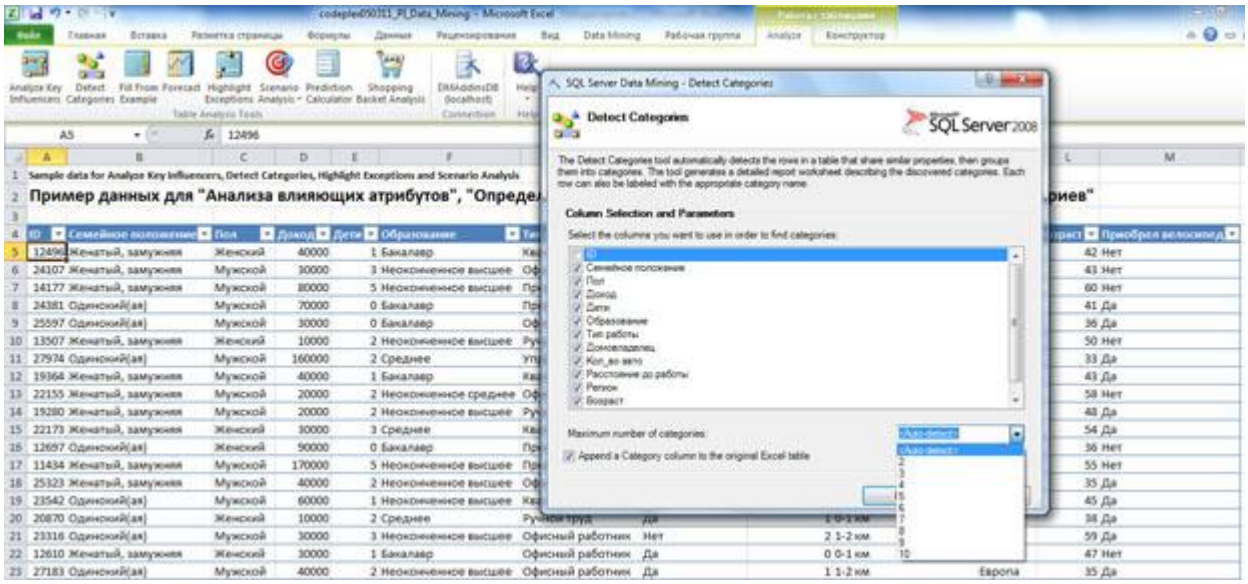
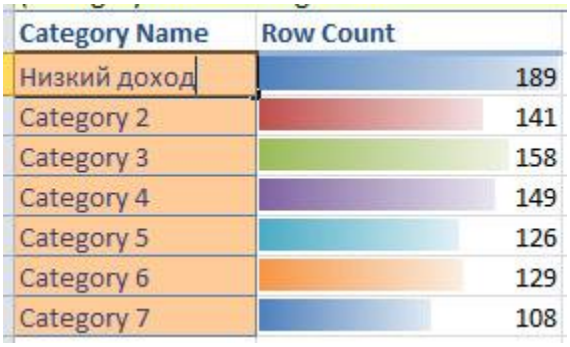


Рис.7. Вибір параметрів, які будуть аналізуватися

Крім вказівки врахованих параметрів, можна явно вказати число категорій (або залишити за замовчуванням автоматичне визначення). Також за замовчуванням поставлений прапорець "Appenda Category column to the original Excel table", який вказує, що до записів у вихідній таблиці буде додано вказівку на категорію.

Сформований звіт містить 3 розділу. У першому вказані певні інструментом категорії і число рядків, що потрапляють в кожну з них (Рис.8). Поле з назвою категорії допускає редагування і можна зіставити категорії більш значуще назву. Наприклад, як буде показано нижче, для клієнтів першої категорії характерний низький дохід і її можна так і назвати. Коли ми

введемо цю назву, всюди крім діаграми Category Profiles Chat, воно автоматично замінить "Category 1" (щоб назва поміняти і на діаграмі, треба натиснути <Alt> + <Ctrl> + <F5>).



Category Name	Row Count
Низкий доход	189
Category 2	141
Category 3	158
Category 4	149
Category 5	126
Category 6	129
Category 7	108

Рис.8. виділені категорії

Наступний розділ звіту описує характеристики виділених категорій і ступінь впливу кожного параметра (Рис.9). Замовчуванням відображається інформація тільки по одній категорії, але клацанням миші по іконці фільтра на заголовку таблиці можна відображати всі канали категорій або якогось їх поєднання, як це показано на рис.

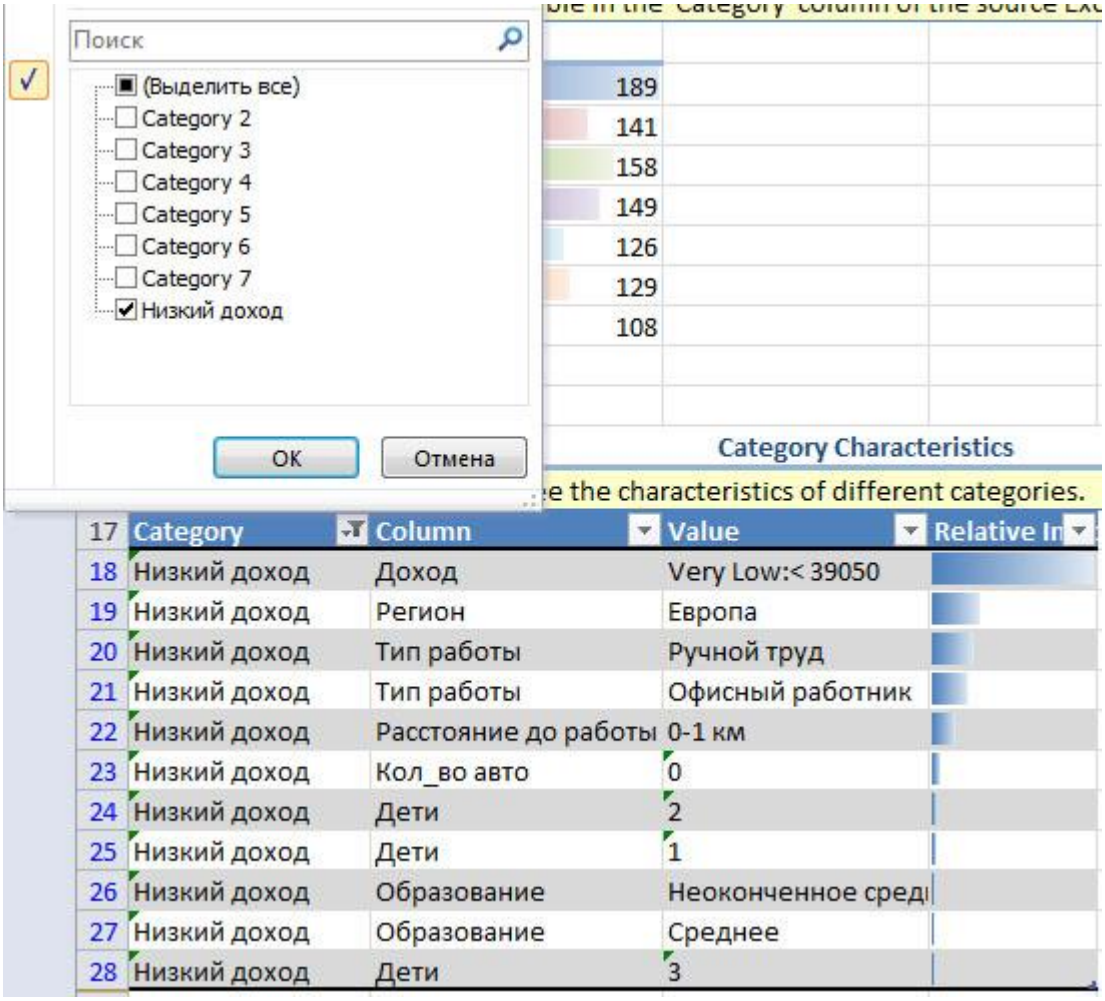


Рис.9. опис категорії

Третій розділ звіту - це діаграма профілів категорій. Вона показує кількість рядків даних в кожній категорії з кожним значенням обраних параметрів. За замовчуванням відображається тільки один параметр. Для розглянутого прикладу це вік. Але в нижній частині діаграми є фільтр Column, за допомогою якого можна змінити число параметрів. Наприклад, на Рис.10 для кожної категорії відображається розподіл за віком і доходу. З нього видно, що клієнти перейменованої нами категорії "Низький дохід" насправді мають дуже низький дохід. А клієнти категорії 3 в переважній більшості дуже молоді.

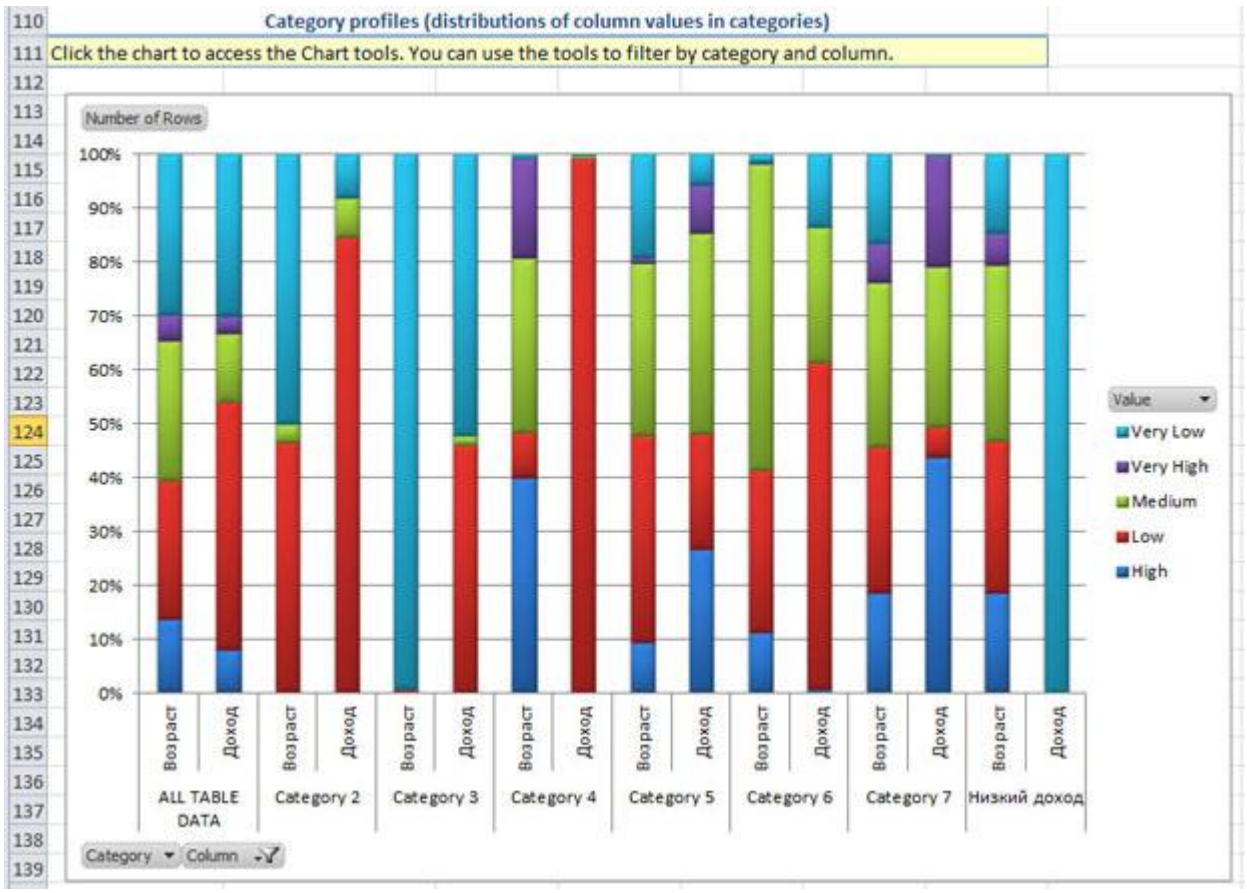


Рис.10. Діаграма профілів категорій

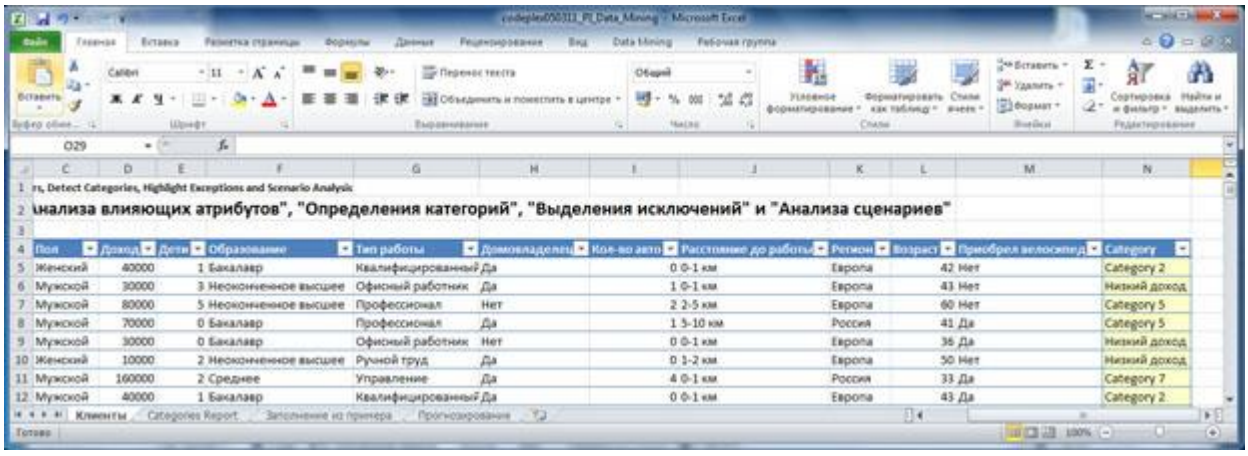


Рис.11. Зіставлення категорій записів у вихідній таблиці

Рис.11 показує, що всіх записів вихідної таблиці тепер сопоставлена категорія, до якої вони належать. А за допомогою фільтрів можна переглянути записи, які стосуються обраної категорії.

Завдання 1. Перейменуйте категорію Category 3.

Завдання 2. Проведіть аналіз параметрів, що характеризують залишилися категорії, і дайте їм осмислені назви.