**Лаб робота 6:**

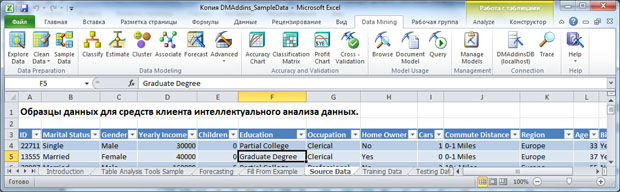
**Использование инструментов Data Mining Client для Excel 2007 для подготовки данных**

**Аннотация:**Данная лабораторная работа описывает возможности инструментов, относящихся к Data Mining Client для Excel 2007, в части подготовки данных для анализа.

**Ключевые слова:**[excel](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#keyword1), [анализ](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#keyword2), [меню](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#keyword4), [лист](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#keyword5), [группа](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#keyword6), [значение](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#keyword7), [диаграмма](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#keyword10), [графика](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#keyword13), [сортировка](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#keyword16), [буфер](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#keyword17), [параметр](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#keyword20), [Гистограмма](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#keyword22), [диапазон](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#keyword27), [опция](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#keyword30), [уровень детализации](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#keyword33), [расстояние](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#keyword34), [mile](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2" \l "keyword35), [процент](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#keyword40), [точное число](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#keyword41), [место](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#keyword42), [таблица](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#keyword45), [выборка](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#keyword46)

Рассмотренные в предыдущих лабораторных работах "Средства анализа таблиц для *Excel*" (**TableAnalysisTools**) позволяют быстро провести "стандартный" *анализ* имеющихся данных. В то же время, этот набор инструментов не предоставляет особых возможностей по подготовке данных к анализу, оценке результатов и т.д. Из *Excel* это можно сделать, используя клиент интеллектуального анализа данных (**DataMiningClient**), который также входит в набор надстроек интеллектуального анализа. В ходе ["Надстройки интеллектуального анализа данных для MicrosoftOffice"](https://www.intuit.ru/studies/courses/2312/612/lecture/13263), отмечалось, что желательно сделать полную установку надстроек, в которую входит и **DataMiningClient**.

Откроем уже использовавшийся нами набор данных, входящий в поставку надстроек (*меню* **"Пуск"**, найдите **Надстройки интеллектуального анализа данных->Образцы данных Excel**).Чтобы можно было спокойно вносить изменения, лучше сохранить его под новым именем.Перейдите на *лист* **"Исходные данные" (SourceData)** и щелкните на закладке **DataMining**. Лента с предлагаемыми инструментами представлена на [рис. 13.1](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.1).

[](https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_01.jpg)

[увеличить изображение](https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_01.jpg)  
**Рис. 13.1.**Инструменты Data Mining Client

Первая *группа* инструментов (**Data Preparation - Подготовка данных**), позволяет провести первое знакомство с набором данных и подготовить его для дальнейшего анализа.

Например, в предыдущих работах мы неоднократно сталкивались с тем, что ряд алгоритмов (MicrosoftNaiveBayes и др.) требуют предварительной дискретизации непрерывных значений числовых параметров. Но в ряде случаев пользователю желательно посмотреть возможные диапазоны, уточнить их число и т.д. Отдельный интерес может представлять и распределение строк по значению выбранного параметра.

### Explore Data

Инструмент **Explore Data** позволяет проанализировать значения столбца (или диапазона ячеек) и отобразить их на диаграмме. Рассмотрим его работу на примере значения годового дохода клиента (**Income**). Дополнительный интерес представляет то, что это *значение* может рассматриваться и как непрерывное, и как дискретное. Итак, запускаем инструмент ([рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2)).

|  |  |
| --- | --- |
| 1 | 2 |
| https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_02-1.jpg | https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_02-2.jpg |
| 3 | 4 |
| Инструмент Explore Data  **Рис. 13.2.**Инструмент Explore Data | https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_02-4.jpg |

В процессе работы потребуется указать, для какой таблицы (или диапазона ячеек) и столбца будет проводиться *анализ* ([рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2)-1 и [рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2)-2). После чего указанные значения будут проанализированы и результат представлен в виде гистограммы.

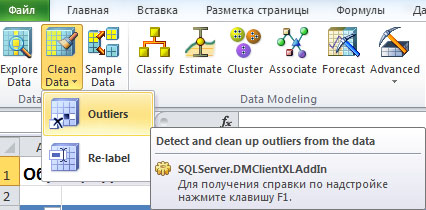
Как уже отмечалось выше, *значение* годового дохода можно рассматривать и как непрерывное, и как дискретное (за счет того, что в нашем наборе данных присутствуют только значения, кратные 10 тысячам). Для непрерывного значения будет предложен вариант разбиения на диапазоны ([рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2)-3). Число диапазонов можно поменять и *диаграмма* с распределением значений будут построена заново. Нажав кнопку **"Add New Column"** можно добавить в исходную таблицу новый столбец с интервалами годового дохода. Например, если для строки *значение* **Yearly Income = 30000**, то *значение* нового параметра **Yearly Income 2** при использовании представленного на рисунке разбиения будет **"'30000 - 50000"** (именно так, с апострофом в начале, чтобы рассматривалось как строковое). В ходе интеллектуального анализа,полученный столбец может использоваться вместо исходного (включение обоих столбцов одновременно нежелательно).

Кнопками с изображениями *графика* и гистограммы (на [рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2)-3, [рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2)-4 они подчеркнуты), можно указать тип анализируемого значения - непрерывное или дискретное. Если *значение* годового дохода рассматриваем как дискретное, то для него будет построена *диаграмма*, показывающая распределение числа строк по значению годового дохода ([рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2)-4). При этом *сортировка* производится по убыванию числа строк с данных значением, из-за чего первый столбец гистограммы соответствует значению **"60000"**, второй - **"40000"** и т.д. Сформированную гистограмму можно скопировать в *буфер* (кнопка правее кнопки **"Add New Column"**, [рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2)-3, [рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2)-4) и использовать для дальнейшей работы.

### Clean Data

Инструмент **Clean Data**([рис. 13.3](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#image.13.3)) позволяет подготовить данные для анализа, отбросив нетипичные или ошибочные данные (выбросы), а также проведя замену отдельных значений. Как отмечается в документации, под выбросом подразумевается *значение* данных, являющееся проблематичным по одной из следующих причин:

* значение находится за пределами ожидаемого диапазона;
* данные были введены неправильно;
* значение отсутствует;
* данные представляют собой пробел или пустую строку;
* значение может значительно отклониться от распределения, которому подчиняются данные в модели.



**Рис. 13.3.**Инструмент CleanData

Использование данного инструмента проиллюстрируем на примере все той же таблицы с данными о клиентах (*лист* **Source Data**). Обратимся к столбцу с возрастом. Пусть нам нужно очистить набор данных от информации о нехарактерных по возрасту покупателях. Запускаем инструмент **Clean Data->Outliers**, в окне аналогичном представленному на [рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2)-1 выбираем таблицу для анализа, затем в окне **Select Column**([рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2)-2)- столбец **Age**.

В рассматриваемом наборе данных есть строки со значениями столбца **Age** от 25 до 96 лет. Если этот *параметр* считаем непрерывным, то он будет представлен графиком, где по оси X указывается возраст, по оси Y-число клиентов с таким возрастом. В наборе данных доля клиентов преклонного возраста очень мала. На [рис. 13.4](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#image.13.4)-1 показано, что установив пороговое *значение* в 75 лет, мы отбрасываем заштрихованный "хвост", включающий нехарактерные значения (покупатели велосипедов в возрасте от 76 до 96 лет, которых подавляющее меньшинство).

Во многом аналогично выглядит работа с параметром, принимающим дискретные значения. Для него строится *гистограмма*, а для определения порога нужно указать минимальное число примеров, "поддерживающих" *значение*. Например, на [рис. 13.4](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#image.13.4)-2, установлено пороговое *значение* в 15. К сожалению, при большом числе столбцов гистограммы,значения параметра на ней не отображаются. Поэтому не понять, что именно попадает в "хвост" распределения.

|  |  |
| --- | --- |
| 1 | 2 |
| https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_04-1.jpg | https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_04-2.jpg |
| 3 | 4 |
| Использование инструмента CleanData для исключения выбросов  **Рис. 13.4.**Использование инструмента CleanData для исключения выбросов | https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_04-4.jpg |

Итак, мы выделили нехарактерные данные. Теперь нужно определить, что с ними делать. Предлагаемые мастером решения несколько отличаются для случаев непрерывного и дискретного параметра. Соответствующую строку можно удалить (**Delete rows containing outliners**) или заменить *значение* параметра на пустое (**Change value to null**). Кроме того, для непрерывных данных ([рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2)-3) можно заменить нехарактерное *значение* средним или граничным (сверху или снизу, в зависимости от того, какой *диапазон* отбрасывается). Для дискретного параметра ([рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2)-4) можно указать *значение* (из числа уже имеющихся в наборе), на которое будут заменяться "выбросы".

Последнее окно мастера (оно на рисунке не представлено) предлагает выбрать, куда заносить изменения - в исходные данные (**Change data inplace**), в их копию на новом листе *Excel* (**Copy sheet data with changes to a new work sheet**)или в новый столбец в исходной таблице (**Add as a new column to the current work sheet**).Последняя *опция* для случая удаления строк недоступна.

### CleanData.Re-label

В некоторых случаях в исходных данных могут быть значения, которые затрудняют автоматизированный *анализ*. Например, есть *параметр* "город" и среди его значений - Санкт-Петербург, С-Петербург, СПб. Для того, чтобы в процессе интеллектуального анализа эти значения учитывались корректно, надо их заменить на одно. Для этого можно использовать инструмент **Re-label**. Его же можно применить, если требуется снизить *уровень детализации* значений параметра. Надо отметить, что инструмент работает только с дискретными значениями (ну или рассматриваемыми как дискретные).

Для примера, в таблице с информацией о клиентах нам надо уменьшить число значений параметра **CommuteDistance** (*расстояние* ежедневных поездок). Исходные значения "0-1 *Miles*", "1-2 *Miles*", "2-5 *Miles*", "5-10 *Miles*", "10+ *Miles*". Пусть все, что меньше 2 миль, будет "близко", остальное - "далеко". Добавим в таблицу две пустые строки и укажем для одной **CommuteDistance** "близко",для другой - "далеко". Делается это потому, что значения, на которые заменяем, тоже должны присутствовать в столбце.

|  |  |
| --- | --- |
| 1 | 2 |
| Замена обозначений  **Рис. 13.5.**Замена обозначений | https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_05-2.jpg |

Запустим инструмент: **CleanData->Re-label**. Первые два экрана, как и ранее, позволяют указать таблицу и столбец. Далее указываем порядок замены ([рис. 13.5](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#image.13.5)-1) и выбираем создание нового столбца ([рис. 13.5](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=2#image.13.5)-2), чтобы не потерять исходные данные. Замена будет произведена, после чего не забудем удалить добавленные пустые строки с "близко"-"далеко".

### SampleData

Последний инструмент в группе **Data Preparation** называется **Sample Data** (Образцы данных). Он позволяет решить задачу формирования обучающего и тестового множеств данных, а также выполнять "балансировку" данных.

В тех случаях, когда используемый метод интеллектуального анализа требует предварительного обучения модели (например, для решения задачи классификации) необходимо сформировать несколько наборов данных - для обучения модели, проверки ее работы, собственно анализа. Инструмент **Sample Data** позволяет подготовить нужные наборы.

Пусть необходимо случайным образом разделить имеющийся набор данных на обучающую и тестовую выборку. Для этого надо запустить инструмент **Sample Data**, указать откуда берем данные для обработки ([рис. 13.6](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.6)-1) и тип формируемой выборки. Сначала сделаем случайную выборку, т.е. тип - **Random Sampling** ([рис. 13.6](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.6)-2). Далее указывается *процент* записей из исходного набора (или *точное число* записей) помещаемых в выборку ([рис. 13.6](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.6)-3) и *место* для сохранения полученных результатов. На [рис. 13.6](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.6)-4 видно, что можно отдельно сохранить сформированную выборку и данные, в нее не попавшие. В итоге можем получить обучающий и тестовый наборы. Хотелось бы обратить внимание на возможность использования внешнего источника данных при формировании выборки ([рис. 13.6](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.6)-1). Это позволяет использовать данные хранящиеся на MS SQLServer для формирования наборов значений. Но как отмечается в описании инструмента, при использовании внешнего источника данных в окне, представленном на [рис. 13.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=1#image.13.2), будет доступен только *параметр* случайной выборки.

При использовании средств интеллектуального анализа для обнаружения редких событий, в обучающем наборе рекомендуется увеличить частоту появления нужного события по сравнению с исходными данными. Формирование подобной выборки часто называют балансировкой данных, и инструмент **SampleData** позволяет ее выполнить.

С помощью инструмента **Explore Data** проанализируем распределение клиентов в наборе данных по регионам. На [рис. 13.7](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.7)-1 видно, что примерно пятая часть клиентов у нас из региона **Pacific** (будем считать это Азиатско-Тихоокеанским регионом). Сформируем набор данных, где таких клиентов будет 50 %.

|  |  |
| --- | --- |
| 1 | 2 |
| https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_06-1.jpg | https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_06-2.jpg |
| 3 | 4 |
| Инструмент Sample Data  **Рис. 13.6.**Инструмент Sample Data | https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_06-4.jpg |

Запустим инструмент **Sample Data**, укажем в качестве источника данных используемую таблицу *Excel* и выберем вариант формирования избыточной выборки с балансировкой данных (**Oversample to balance data distributions**, [рис. 13.7](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.7)-2). Далее укажем столбец, для которого выполняется балансировка, и частоту появления нужного значения и размер выборки ([рис. 13.7](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.7)-3).Будет создана новая *таблица* с указанным пользователем названием. Снова применим **Explore Data** и убедимся в том, что *выборка* сформирована в соответствии с указанными выше требованиями ([рис. 13.7](https://www.intuit.ru/studies/courses/2312/612/lecture/13273?page=3#image.13.7)-4).

|  |  |
| --- | --- |
| 1 | 2 |
| https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_07-1.jpg | https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_07-2.jpg |
| 3 | 4 |
| Формирование выборки с заданным распределением клиентов по регионам  **Рис. 13.7.**Формирование выборки с заданным распределением клиентов по регионам | https://www.intuit.ru/EDI/14_01_19_2/1547417979-9274/tutorial/1036/objects/13/files/22_07-4.jpg |

**Задание.** Проведите описанную в лабораторнойобработку выбранного набора данных.