**Самостоятельная работа 3:**

**Использование инструментов "HighlightExceptions" и "ScenarioAnalysis"**

**Аннотация:**Лабораторная работа посвящена использованию инструментов "Выделение исключений" ("HighlightExceptions") и "Анализ сценариев" ("ScenarioAnalysis").

**Ключевые слова:**[excel](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1" \l "keyword1), [анализ исключений](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#keyword2), [ПО](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#keyword3), [поле](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#keyword4), [exception](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#keyword6), [threshold](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#keyword7), [значение](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#keyword8), [лист](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#keyword10), [расстояние](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#keyword15), [Раскраска](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#keyword18), [таблица](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#keyword19), [анализ](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#keyword21), [запуск](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#keyword23), [выборка](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=2#keyword28)

В качестве учебного набора данных, как и в прошлых лабораторных, будем использовать локализованный пример для *Excel*, взятый с <http://russiandmaddins.codeplex.com/>

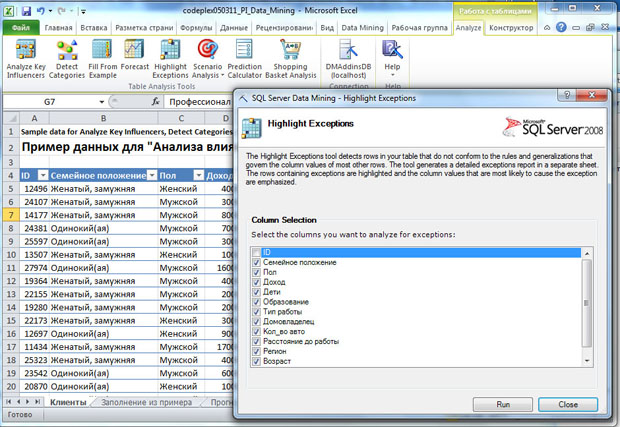
### Выделение исключений

Как следует из названия, инструмент позволяет выявить данные, выделяющиеся среди имеющегося набора. Это может быть полезно в ряде случаев. Во-первых, это могут быть ошибочные данные (например, результаты ошибки оператора при вводе каких-то значений). Во-вторых, исключения могут представлять отдельный интерес (как, например, в случае обнаружения мошеннических действий с банковскими картами и т.п.). Кроме того, *анализ исключений* может рассматриваться как предварительная часть интеллектуального анализа данных с помощью других методов. В частности, это позволяет исключить попадание нетипичных примеров в обучающую выборку.

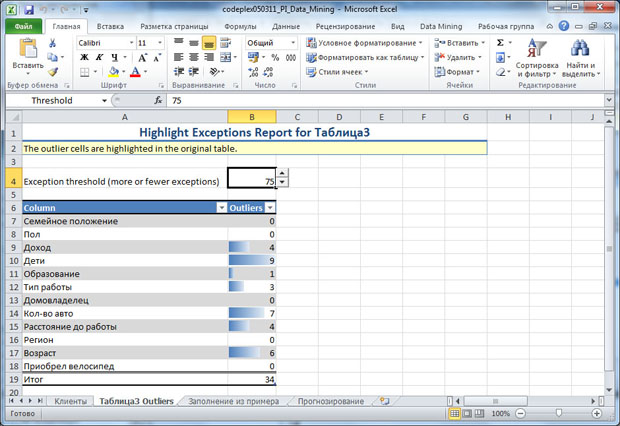
В ходе работы инструмент **HighlightExceptions** создает временную модель интеллектуального анализа с использованием алгоритма MicrosoftClustering. Для каждой анализируемой строки оценивается степень принадлежности выявленным кластерам.Значения, находящиеся далеко от всех кластеров, помечаются как исключения.

При запуске инструмента можно отметить столбцы, не учитываемые при анализе. В рекомендациях *по* использованию [[1](https://www.intuit.ru/studies/courses/2312/612/literature#literature.1),[3](https://www.intuit.ru/studies/courses/2312/612/literature#literature.3)] указывается, что желательно исключить из анализа столбцы с уникальными значениями (имена, идентификаторы), а также содержащие много пустых значений или произвольный текст. На [рис. 7.1](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#image.7.1) видно, что при анализе набора данных **"Клиенты"**инструмент предлагает исключить из рассмотрения *поле* **ID**.

*По* итогам работы (а работает этот инструмент несколько дольше рассмотренных нами ранее) формируется отчет ([рис. 7.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#image.7.2)) и в исходном наборе данных исключения выделяются цветом ([рис. 7.3](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#image.7.3)).

[](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_01.jpg)

[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_01.jpg)  
**Рис. 7.1.**Запуск инструмента HighlightExceptions

[](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_02.jpg)

[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_02.jpg)  
**Рис. 7.2.**Отчет по проведенному анализу данных

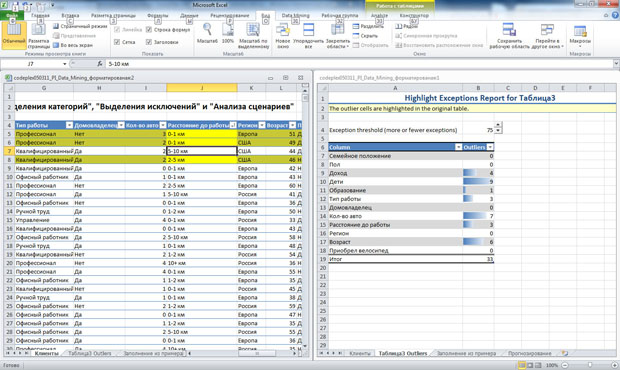
[](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_03.jpg)

[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_03.jpg)  
**Рис. 7.3.**Исключения выделяются цветом, что позволяет произвести сортировку

На [рис. 7.2](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#image.7.2) видно, что инструмент позволяет указать порог отклонения от нормы (*Exception* *threshold*), измеряемый в процентах (оценка вероятности того, что выделенное *значение* относится к исключениям). Уменьшение порога приведет к тому, что больше записей будет рассматриваться как исключения, увеличение - наоборот. При значении *по* умолчанию в 75 % нашем наборе данных обнаружено 34 исключения. Отчет показывает, в каких столбцах сколько исключений было обнаружено.

Перейдем на *лист* *Excel* с данными. Рассматриваемые как выбросы значения выделяются в таблице цветом: вся строка-коричневым, конкретное *значение* - желтым. Чтобы сгруппировать нужные строки можно воспользоваться функциями *Excel*, позволяющими провести сортировку *по* цвету.

Также можно воспользоваться инструментами вкладки **"Вид"**, чтобы создать новое окно и расположить рядом с окном с отчетом и данными ([рис. 7.4](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#image.7.4)). Пусть в отобранном наборе записей мы обнаружили ошибку.Скажем *расстояние* до работы у некоего клиента из США, обладающего двумя машинами, не "0-1 км", а "5-10 км" (именно поэтому ему нужно в семье 2 машины). Если мы изменим *значение*, будет произведен автоматический пересчет. В случае, представленном на [рис. 7.4](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#image.7.4), новое *значение* уже не рассматривается как выброс.

[](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_04.jpg)

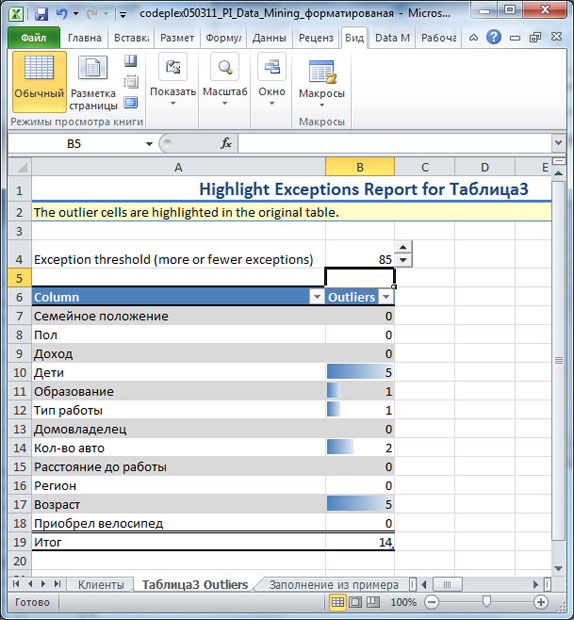
[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_04.jpg)  
**Рис. 7.4.**Исправление ошибочного значения

Обратите внимание, что не только изменилась *раскраска* строки таблицы, но и произошли изменения в отчете, показывающем теперь наличие 33 исключений. Автоматический пересчет работает только в том случае, если сессия работы с аналитическими службами SQLServer остается открытой. Если *таблица* *Excel* была закрыта и снова открыта, то автоматического пересчета не будет (нужно снова провести *анализ*).

Также в описаниях отмечается, что инструмент реагирует только на изменения данных в диапазоне ячеек, использовавшемся при обучении. Если после начала работы инструмента в конец таблицы добавить новые строки, они оцениваться не будут.

Как уже отмечалось выше, если нужно рассматривать только наиболее сильные выбросы, можно увеличить *значение* порога отклонения и инструмент изменит оценки в соответствии с заданным значением ([рис. 7.5](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#image.7.5)).

Повторный *запуск* инструмента удалит результаты предыдущего анализа. Учитывая,что проводимые инструментом изменения достаточно сложны (*раскраска* строк таблицы и т.д.), если нужно удалить результаты работы, рекомендуется запустить повторный *анализ*, согласиться с удалением результатов и потом в окне, аналогичном представленному на [рис. 7.1](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=1#image.7.1), нажать кнопку **Close**(отказаться от анализа данных).

[](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_05.jpg)

[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_05.jpg)  
**Рис. 7.5.**Увеличение порога отклонения уменьшает число исключений

**Задание.** Проведите *анализ исключений* для набора данных **"Клиенты"** и значения порога в 90%. Предложите интерпретацию полученных результатов.

**Задание 2.** Проведите *анализ исключений* для набора данных **"Прогнозирование"** (продажи по месяцам в разных регионах). Предложите интерпретацию полученных результатов.

### Анализ сценариев

Инструмент **Scenario Analysis** позволяет моделировать влияние, оказываемое изменением одного из параметров (значений одного столбца) на другой, связанный с первым. В основе работы инструмента лежит использование алгоритма **Microsoft Logistic Regression**. Для формирования временной модели требуется обучающая *выборка*, содержащая не менее 50 записей [[3](https://www.intuit.ru/studies/courses/2312/612/literature#literature.3)].

Инструмент **Scenario Analysis** включает две составные части - **"Анализ сценария поиска решений" (GoalSeek)** и **"Анализ возможных вариантов"("What-If")**.

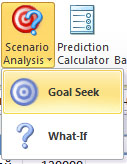
#### "Анализ сценария поиска решений" (GoalSeek)

Использование инструмента **GoalSeek** позволяет оценить, сможем ли мы достичь желаемого значения в целевом столбце, меняя значения выбранного параметра. Инструмент позволяет провести анализ как для одной записи, так и для всей таблицы.

Используя этот инструмент надо быть готовым, что не для всех вариантов запроса может быть получен ответ. Это может быть связано с тем, что в исходных данных нет интересующих нас сочетаний. Также могут быть проблемы из-за типов данных.

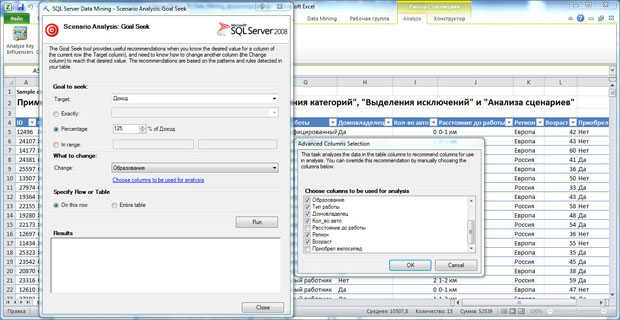
Кроме того, нельзя забывать, что запрос нужно формировать с учетом знаний о предметной области. Например, можно запросить систему, если человек хочет увеличить годовой доход на 20 процентов, надо ли ему приобретать велосипед. И даже получить какой-то ответ. Но понятно, что в такой постановке сам вопрос является бессмысленным.

Пусть мы хотим узнать, как будет влиять образование на уровень достатка человека. Сначала проведем анализ для одной записи. Например, нас интересует клиент с идентификатором **12496** (первая запись в наборе данных).Откройте набор данных **"Клиенты"** и на вкладке **Analysis** выберите **ScenarioAnalysis->GoalSeek** ([рис. 7.6](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=2#image.7.6)).



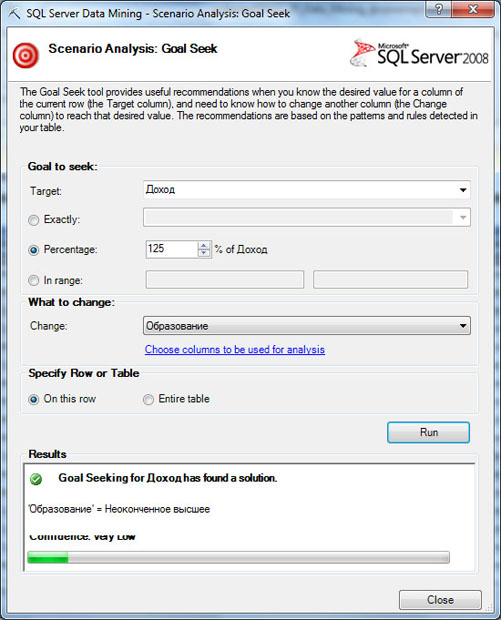
**Рис. 7.6.**Выбор инструмента GoalSeek

В окне параметров ([рис. 7.7](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=2#image.7.7)) укажем целевой столбец - **"Доход"**, а также желаемое значение - **125%** от текущего. В этом случае инструмент считает успешным результат, который не меньше заданного (в нашем примере 40000 \times 1,25 = 50000 и более). Если задаваемое значение меньше 100%, то успешным считается результат, который не больше заданного. Также можно указать точное значение и диапазон (выбрав **"Inrange"**). Для значений, не являющихся числовыми, варианты **"Percentage"** и **"Inrange"** будут неактивны. Для достижения искомого значения будем менять столбец **"Образование"**.

[](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_07.jpg)

[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_07.jpg)  
**Рис. 7.7.**Настройка параметров для GoalSeek

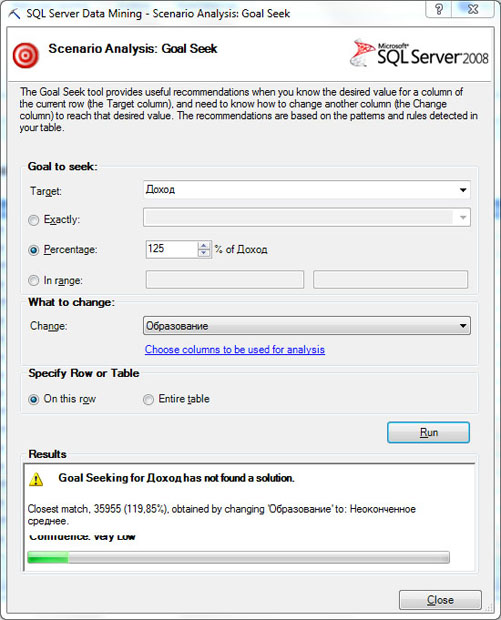
Перейдя по ссылке **"Choose columns to be used for analysis"**, отметим, что при анализе в рассмотрение не берем столбцы **"ID"**, **"Дети"**, **"Расстояние до работы"**, **"Приобрел велосипед"**.После закрытия окна **"Advanced Columns Selection"** стоит еще раз проверить настройки в секции **"Goaltoseek"** - иногда при переходе между окнами переключатель между **"Exactly"**, **"Percentage"** и **"Inrange"** сбрасываетсяв значение по умолчанию (**"Exactly"**)

[](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_08.jpg)

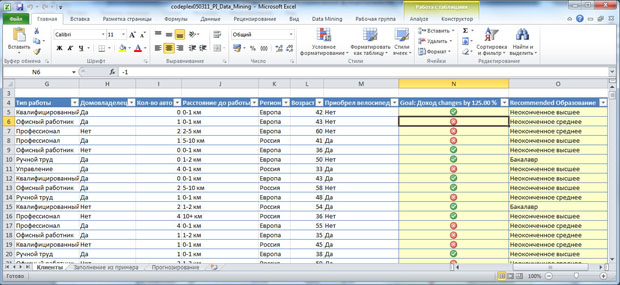
[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_08.jpg)  
**Рис. 7.8.**Результат анализа для одной строки - решение найдено

Результат анализа, выполненного по нажатию кнопки **Run**, представлен на [рис. 7.8](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=2#image.7.8). Для выбранной строки найден шаблон, рекомендующий для параметра **"Образование"** значение **"Неоконченное высшее"**. При этом уровень достоверности - **Confidence**(иногда верхняя часть надписи затирается, как на рисунке), оценивается как очень низкий (**"Very low"**).

Если прейти на следующую строку и снова нажать **Run**, получим результат для новых данных ([рис. 7.9](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=2#image.7.9)). В этом случае, подходящего решения не было найдено, и был предложен наиболее близкий вариант.

[](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_09.jpg)

[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_09.jpg)  
**Рис. 7.9.**Результат анализа для одной строки - решение не найдено

[](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_10.jpg)

[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_10.jpg)  
**Рис. 7.10.**Анализ для всей таблицы

А если в секции **"Specify Rowor Table"** установить переключатель в **"Entire table"**, то сценарии будут посчитаны для всех строк ([рис. 7.10](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=2#image.7.10)). Результаты будут указаны в двух столбцах, добавленных в исходную таблицу. Для тех строк, которые отмечены крестиком в красном круге, соответствующего желаемому сценарию шаблона найдено не было.

**Задание.** Проведите анализ для отдельной строки и таблицы, аналогичный описанному выше. Прокомментируйте результаты.

**Примечание.** Запуск процедуры анализа для ряда других комбинаций столбцов (например - целевой столбец "покупка велосипеда" = "да", независимая переменная - "расстояние до работы") приводит к ошибке **"Query (1, 50) Синтаксический анализатор: Неверный синтаксис "value"."**, видимо связанной с некорректной обработкой некоторых типов данных.

#### "Анализ возможных вариантов"("What-If")

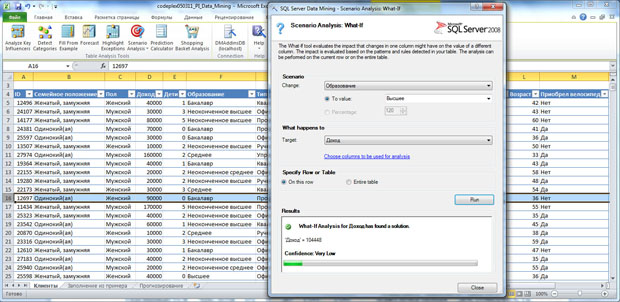
Инструмент **What-If** позволяет решить обратную по отношению к **GoalSeek** задачу: оценить значение целевой переменной при определенном изменении заданного параметра.

Например, можно оценить, как изменился бы уровень дохода человека, если бы повысился его уровень образования. Перейдем на запись с идентификатором **12697** и запустим инструмент: **Scenario Analysis->What-If**. Укажем параметры сценария: образование меняется на **"Высшее"** и целевой столбец **"Доход"**. Полученный для строки результат показывает, что при изменении уровня образования доход может несколько вырасти (исходное значение 90000, среднее значение для нового шаблона 104448). Но степень уверенности в прогнозе не слишком высокая.

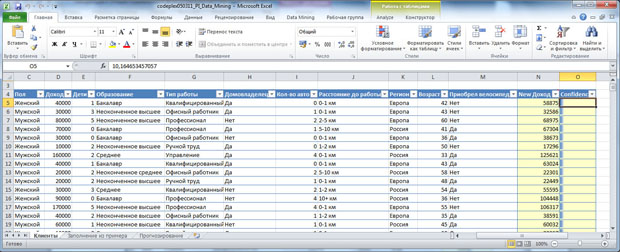
Аналогично предыдущему инструменту,подобный анализ сценария можно сделать и для всей таблицы целиком. В этом случае к исходной таблице добавляются два столбца - один показывает новое значение целевого параметра, второй - оценку достоверности ([рис. 7.12](https://www.intuit.ru/studies/courses/2312/612/lecture/13266?page=2#image.7.12)). Достоверность оценивается числом от 0 до 100: 100 - максимальная достоверность (абсолютная уверенность в прогнозе), 0 - минимальная.

**Задание.** Проведите анализ данных, аналогичный описанному выше.

Для того чтобы удалить результаты работы с таблицей инструментов **What-If** и **Scenario Analysis**, достаточно удалить добавленные столбцы. При работе с отдельными строками, никаких дополнительных действий не требуется.

[](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_11.jpg)

[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_11.jpg)  
**Рис. 7.11.**Работа инструмента What-Ifдля отдельной строки

[](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_12.jpg)

[увеличить изображение](https://www.intuit.ru/EDI/03_01_17_9/1483395677-6027/tutorial/1036/objects/7/files/20_12.jpg)  
**Рис. 7.12.**Прогноз What-Ifдля всей таблицы