Дубликаты и уникальные

Приходилось ли вам когда-нибудь сравнивать между собой два больших списка, где часть элементов совпадает, а часть – нет? Или искать повторения в большой таблице? Этот раздел посвящен различным аспектам работы с таблицами, где некоторые значения могут встречаться больше одного раза.

В этой главе мы научимся:

- **Находить, выделять цветом и удалять**, при необходимости, **дубликаты** в большом списке.
- Подсчитывать **количество уникальных** элементов в большом списке с повторениями.
- **Извлекать** только **неповторяющиеся элементы** из списка.

Подсчет количества уникальных значений в диапазоне

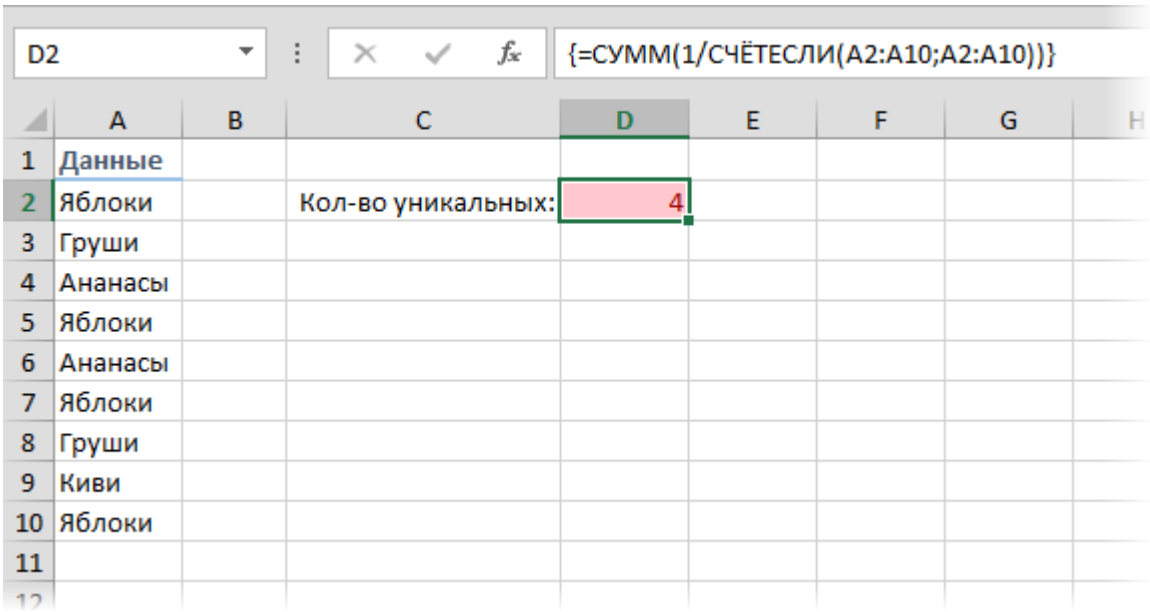
Предположим, что у нас есть диапазон с данными, в котором некоторые значения повторяются больше одного раза:

	A	B
1	Данные	
2	Яблоки	
3	Груши	
4	Киви	
5	Яблоки	
6	Ананасы	
7	Яблоки	
8	Груши	
9	Киви	
10	Яблоки	
11		

Задача в том, чтобы подсчитать количество уникальных (неповторяющихся) значений в диапазоне. В приведенном выше примере, как легко заметить, на самом деле упоминаются всего четыре товара. Рассмотрим несколько способов решения такой задачи.

Способ 1. Если нет пустых ячеек

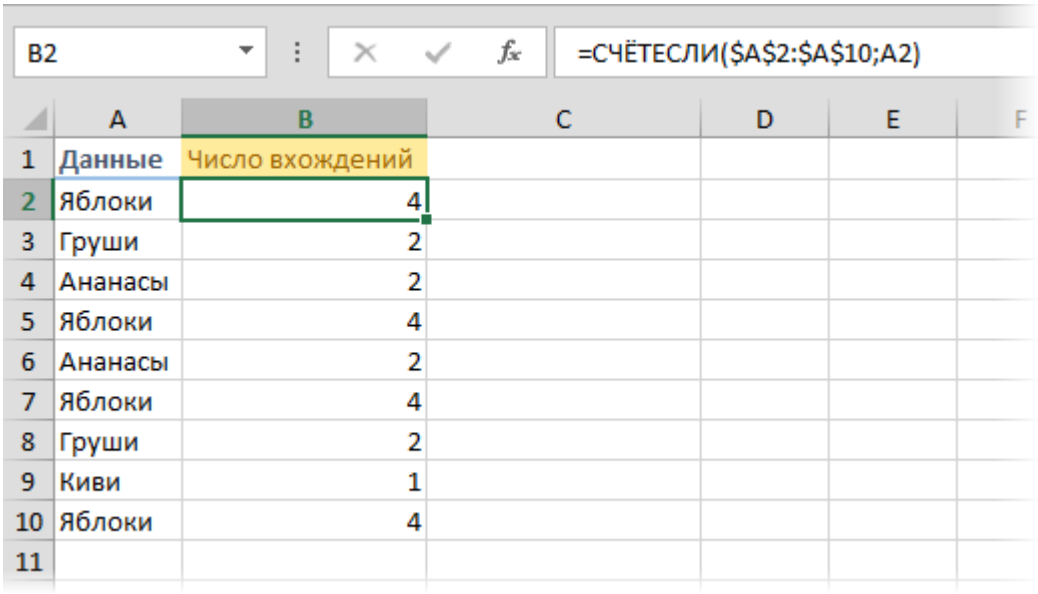
Если вы уверены, что в исходном диапазоне данных нет пустых ячеек, то можно использовать короткую и элегантную формулу массива:



В английской версии это будет выглядеть как =SUM(1/COUNTIF(A2:A10;A2:A10))

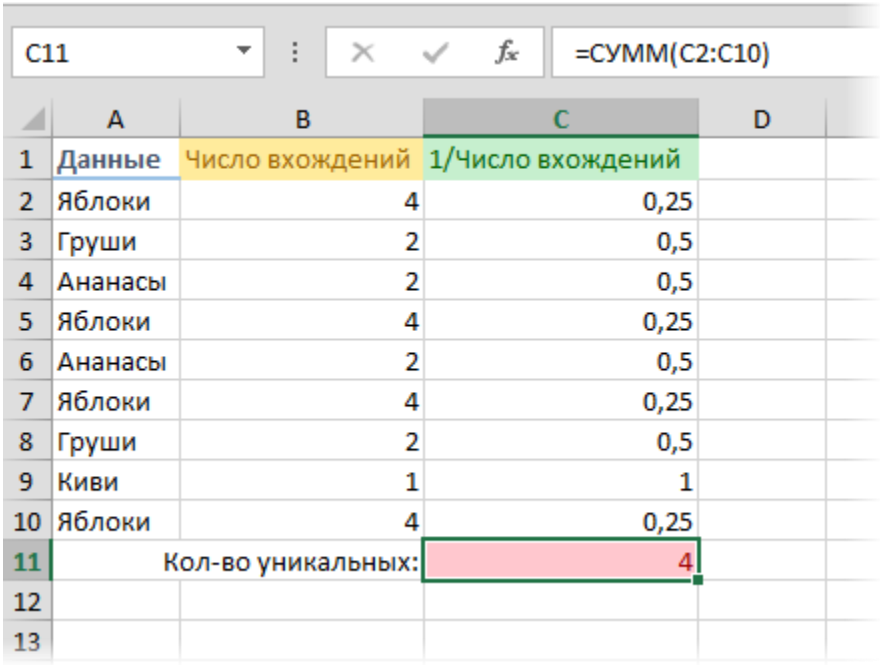
Не забудьте ввести ее как формулу массива, т.е. нажать после ввода формулы не **Enter**, а сочетание **Ctrl+Shift+Enter**.

Технически, эта формула пробегает по всем ячейкам массива и вычисляет для каждого элемента количество его вхождений в диапазон с помощью функции **СЧЕТЕСЛИ (COUNTIF)**. Если представить это в виде дополнительного столбца, то выглядело бы оно так:



	A	B	C	D	E
1	Данные	Число вхождений			
2	Яблоки	4			
3	Груши	2			
4	Ананасы	2			
5	Яблоки	4			
6	Ананасы	2			
7	Яблоки	4			
8	Груши	2			
9	Киви	1			
10	Яблоки	4			
11					

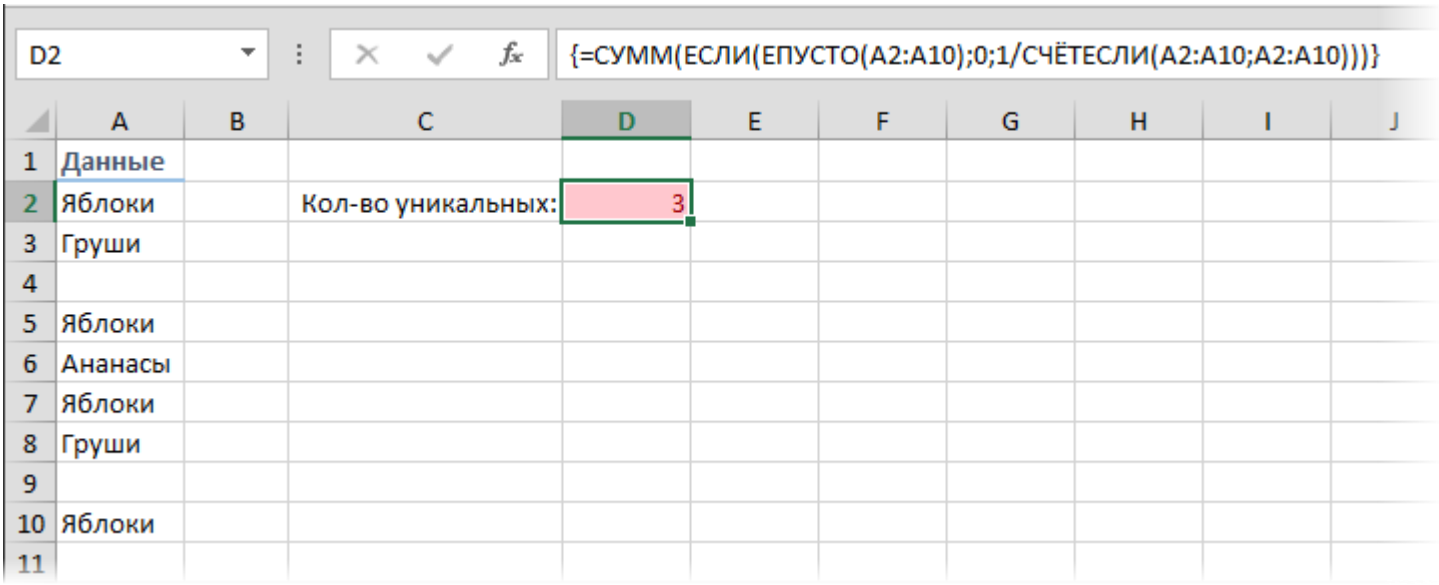
Потом вычисляются дроби **1/Число вхождений** для каждого элемента и все они суммируются, что и даст нам количество уникальных элементов:



	A	B	C	D
1	Данные	Число вхождений	1/Число вхождений	
2	Яблоки	4	0,25	
3	Груши	2	0,5	
4	Ананасы	2	0,5	
5	Яблоки	4	0,25	
6	Ананасы	2	0,5	
7	Яблоки	4	0,25	
8	Груши	2	0,5	
9	Киви	1	1	
10	Яблоки	4	0,25	
11		Кол-во уникальных:	4	
12				
13				

Способ 2. Если есть пустые ячейки

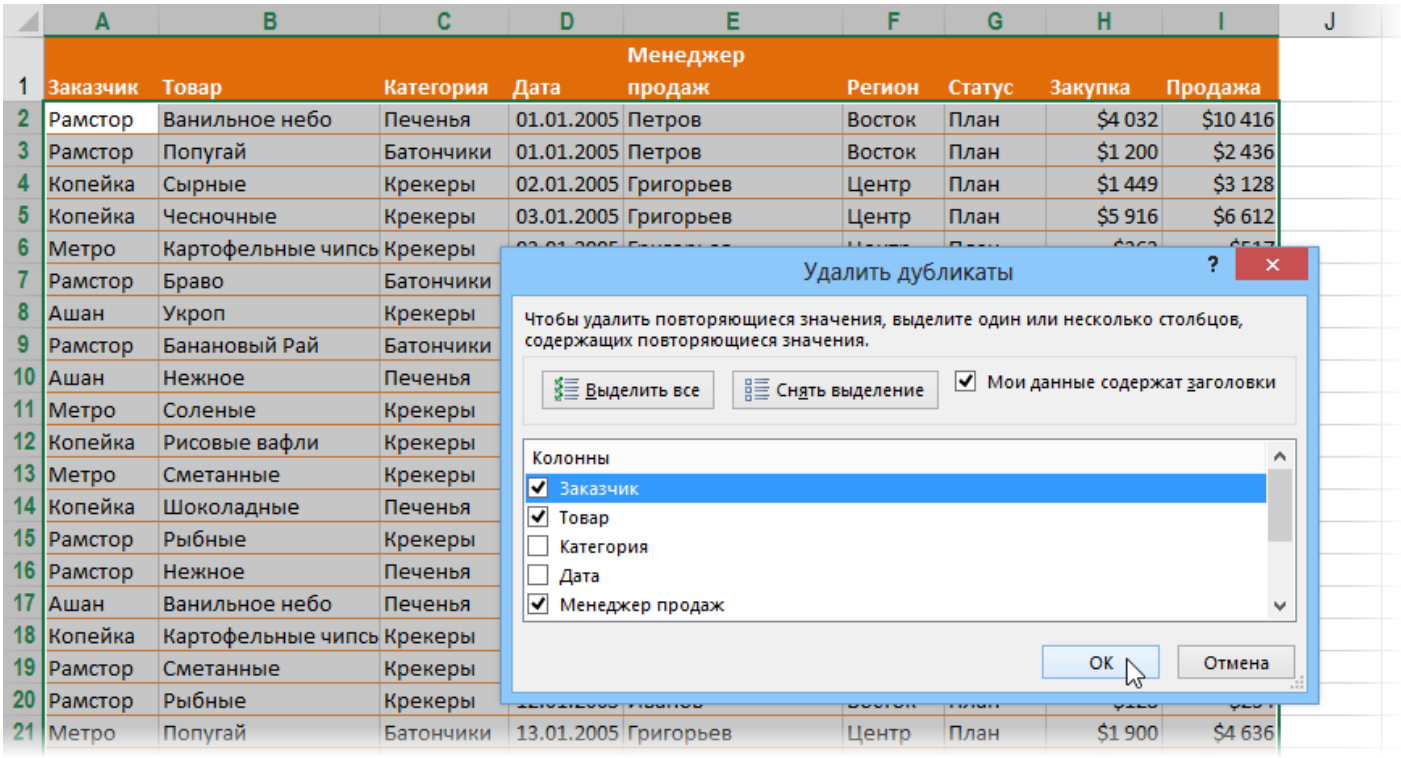
Если в диапазоне встречаются пустые ячейки, то придется немного усовершенствовать нашу формулу массива, добавив проверку на пустые ячейки (иначе получим ошибку деления на 0 в дроби):


D2
{=СУММ(ЕСЛИ(ЕПУСТО(A2:A10);0;1/СЧЁТЕСЛИ(A2:A10;A2:A10)))}
A B C D E F G H I J
1 Данные
2 Яблоки Кол-во уникальных: 3
3 Груши
4
5 Яблоки
6 Ананасы
7 Яблоки
8 Груши
9
10 Яблоки
11

Удаление дубликатов строк

Начиная с 2007-й версии Excel функция удаления дубликатов является стандартной – найти ее можно на вкладке **Данные – Удаление дубликатов (Data – Remove Duplicates)**, установив предварительно активную ячейку в таблицу или выделив ее.

В открывшемся окне нужно с помощью флажков задать те столбцы, по которым необходимо обеспечивать уникальность:



	A	B	C	D	E	F	G	H	I
1	Заказчик	Товар	Категория	Дата	Менеджер продаж	Регион	Статус	Закупка	Продажа
2	Рамстор	Ванильное небо	Печенья	01.01.2005	Петров	Восток	План	$4 032	$10 416
3	Рамстор	Попугай	Батончики	01.01.2005	Петров	Восток	План	$1 200	$2 436
4	Копейка	Сырные	Крекеры	02.01.2005	Григорьев	Центр	План	$1 449	$3 128
5	Копейка	Чесночные	Крекеры	03.01.2005	Григорьев	Центр	План	$5 916	$6 612
6	Метро	Картофельные чипсы	Крекеры						
7	Рамстор	Браво	Батончики						
8	Ашан	Укроп	Крекеры						
9	Рамстор	Банановый Рай	Батончики						
10	Ашан	Нежное	Печенья						
11	Метро	Соленые	Крекеры						
12	Копейка	Рисовые вафли	Крекеры						
13	Метро	Сметанные	Крекеры						
14	Копейка	Шоколадные	Печенья						
15	Рамстор	Рыбные	Крекеры						
16	Рамстор	Нежное	Печенья						
17	Ашан	Ванильное небо	Печенья						
18	Копейка	Картофельные чипсы	Крекеры						
19	Рамстор	Сметанные	Крекеры						
20	Рамстор	Рыбные	Крекеры						
21	Метро	Попугай	Батончики	13.01.2005	Григорьев	Центр	План	$1 900	$4 636

Т.е. если включить все флажки, то будут удалены только полностью совпадающие строки. Если включить только флажок **Заказчик**, то останется только по одной строке для каждого заказчика. Если включить флажки **Заказчик** и **Товар**, то мы увидим все неповторяющиеся комбинации заказчиков и товаров и т.д.

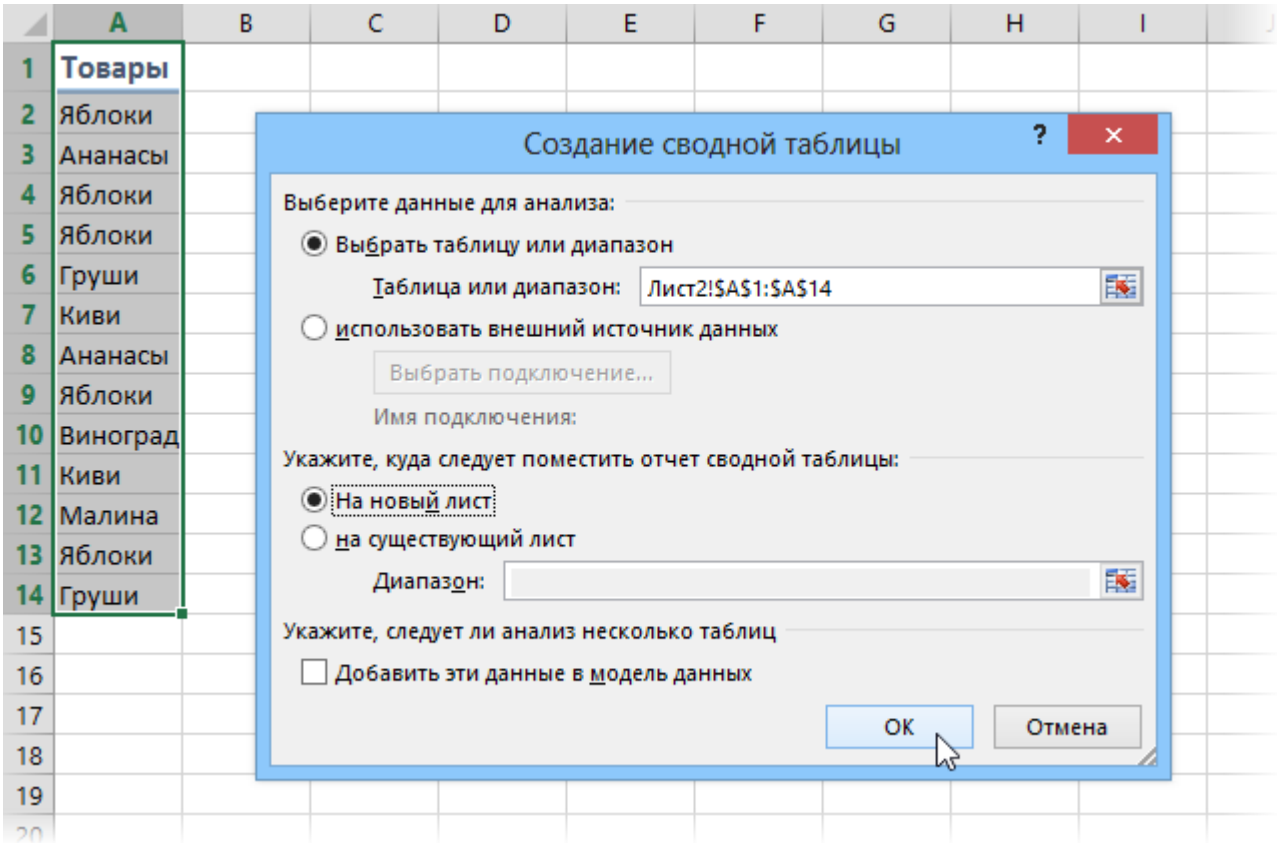
Минусы данного инструмента:

- нет предварительного просмотра результатов – Excel просто молча удалит строки с дубликатами (а иногда хотелось бы сначала на них посмотреть до удаления),
- таблица должна быть с "правильной" однострочной, а не многоэтажной "шапкой" (строкой заголовка).

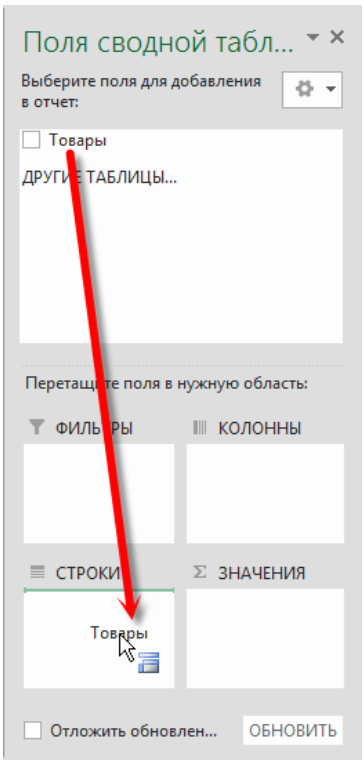
Извлечение списка уникальных элементов из диапазона

Способ 1. Сводная таблица

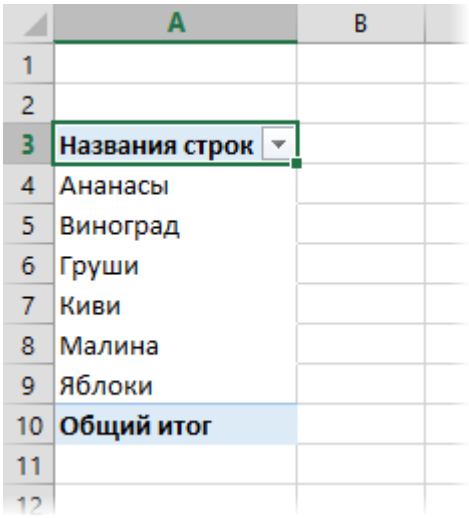
Выделите исходный диапазон или установите в него активную ячейку и выберите на вкладке **Вставка – Сводная таблица (Insert – Pivot Table)**. В открывшемся промежуточном окне жмем **ОК**:



Затем справа, в панели полей сводной таблицы перетаскиваем мышью поле **Товары** в область строк:



Поскольку в области строк сводной таблицы повторов не бывает, мы получим список неповторяющихся элементов из нашего списка, который потом можно скопировать и использовать где угодно:



	A	B
1		
2		
3	**Названия строк**	
4	Ананасы	
5	Виноград	
6	Груши	
7	Киви	
8	Малина	
9	Яблоки	
10	**Общий итог**	
11		
12		

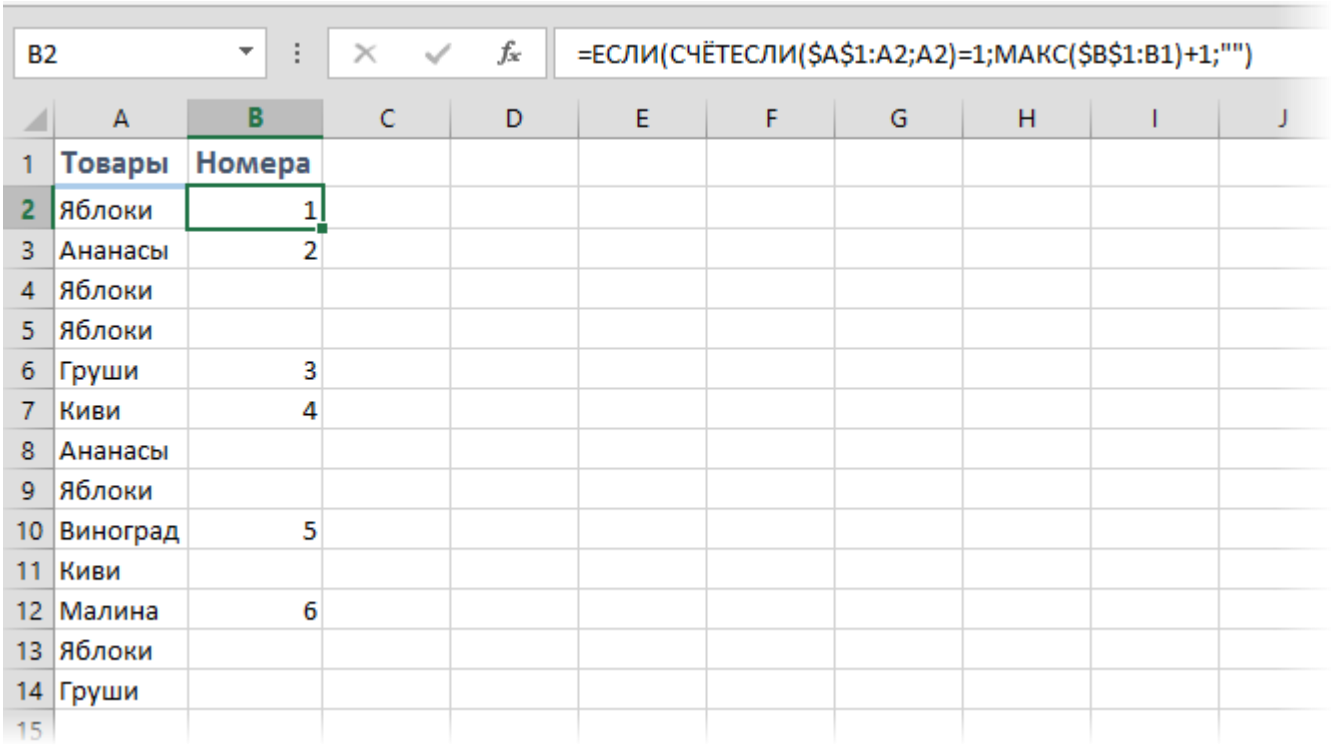
Способ 2. Формулой

Чуть более сложный способ, чем предыдущий, но зато – динамический, т.е. с автоматическим пересчетом. Таким образом, если список редактируется или в него дописываются новые элементы, то они автоматически проверяются на уникальность и отбираются "на лету".

Итак, снова имеем список беспорядочно повторяющихся элементов. Например, такой:

	A	B
1	**Товары**	
2	Яблоки	
3	Ананасы	
4	Яблоки	
5	Яблоки	
6	Груши	
7	Киви	
8	Ананасы	
9	Яблоки	
10	Виноград	
11	Киви	
12	Малина	
13	Яблоки	
14	Груши	
15		

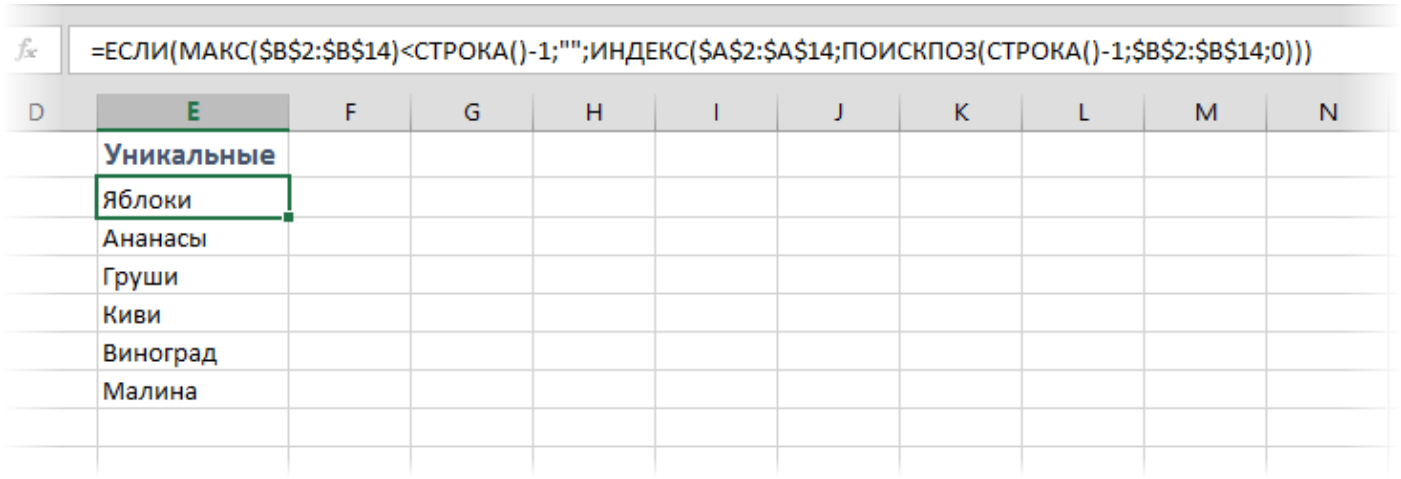
Добавим к нему справа еще один столбец, в котором пронумеруем первые вхождения каждого элемента с помощью вот такой формулы:



	A	B	C	D	E	F	G	H	I	J
1	Товары	Номера								
2	Яблоки	1								
3	Ананасы	2								
4	Яблоки									
5	Яблоки									
6	Груши	3								
7	Киви	4								
8	Ананасы									
9	Яблоки									
10	Виноград	5								
11	Киви									
12	Малина	6								
13	Яблоки									
14	Груши									
15										

Обратите внимание на соответствующее закрепление областей $A$1:A2 и $B$1:B1, где первая ячейка является абсолютной, а вторая – нет, что приводит к "растягиванию" этих диапазонов при копировании формулы.

Теперь осталось извлечь из списка только те товары, напротив которых стоят цифры. Это можно сделать с помощью вот такой формулы:



D	E	F	G	H	I	J	K	L	M	N
	Уникальные									
	Яблоки									
	Ананасы									
	Груши									
	Киви									
	Виноград									
	Малина									

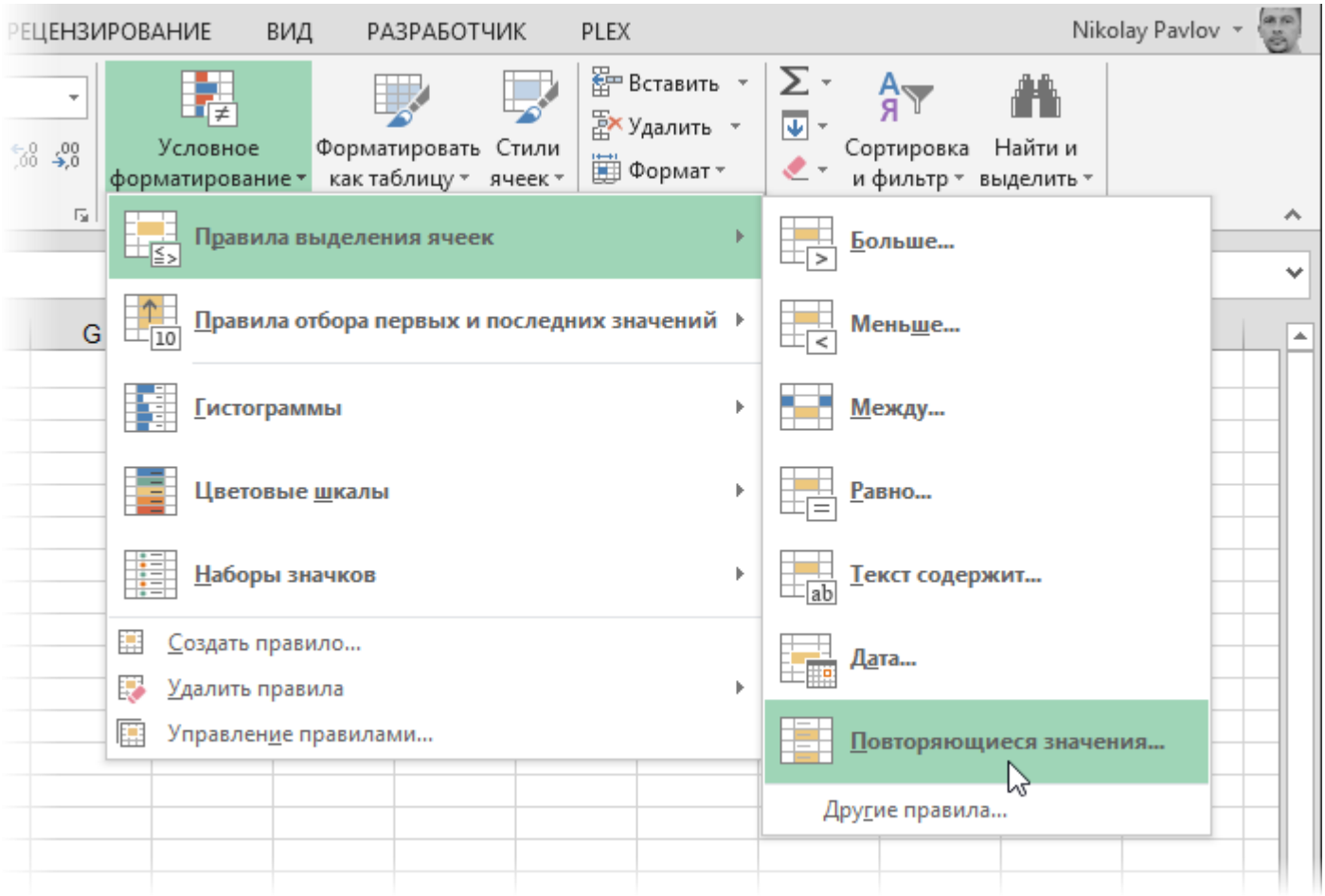
Выделение дубликатов цветом

В одном столбце

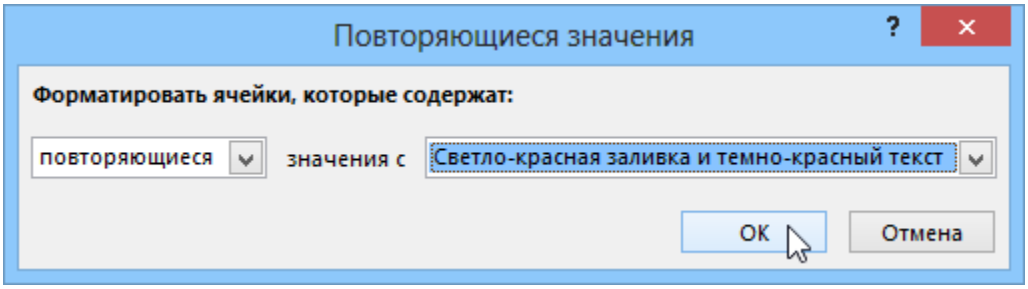
Допустим, что у нас имеется длинный список чего-либо (например, товаров), и мы предполагаем, что некоторые элементы этого списка повторяются более одного раза. Хотелось бы видеть эти повторы явно, т.е. подсветить дублирующие ячейки цветом, например так:

	A
1	**Товары**
2	Queso Cabrales
3	Singaporean Hokkien Fried Mee
4	Mozzarella di Giovanni
5	Manjimup Dried Apples
6	Tofu
7	Manjimup Dried Apples
8	Jack's New England Clam Chowder
9	Louisiana Fiery Hot Pepper Sauce
10	Louisiana Fiery Hot Pepper Sauce
11	Gustaf's Knäckebröd
12	Ravioli Angelo
13	Geitost
14	Sir Rodney's Marmalade

В последних версиях Microsoft Excel реализовать такое очень просто. Выделяем все ячейки с данными и на вкладке **Главная (Home)** жмем кнопку **Условное форматирование (Conditional Formatting)**. Затем выбираем **Правила выделения ячеек – Повторяющиеся значения (Highlight Cell Rules – Duplicate Values)**:



В появившемся затем окне можно задать желаемое форматирование (заливку, цвет шрифта и т.д.) для ячеек с повторами:



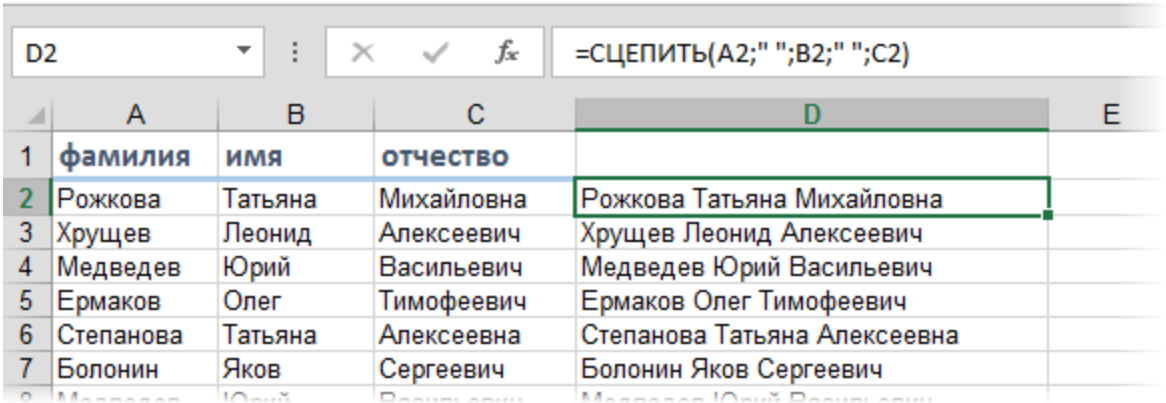
В нескольких столбцах

Усложним задачу. Допустим, нам нужно искать и подсвечивать повторы не по одному столбцу, а по нескольким. Например, имеется вот такая таблица с ФИО в трех колонках:

	A	B	C
1	**фамилия**	**имя**	**отчество**
2	Рожкова	Татьяна	Михайловна
3	Хрущев	Леонид	Алексеевич
4	Медведев	Юрий	Васильевич
5	Ермаков	Олег	Тимофеевич
6	Степанова	Татьяна	Алексеевна
7	Болонин	Яков	Сергеевич
8	Медведев	Юрий	Васильевич
9	Кротова	Анна	Григорьевна

Задача все та же – подсветить совпадающие ФИО, имея в виду совпадение сразу по всем трем столбцам – имени, фамилии и отчества одновременно.

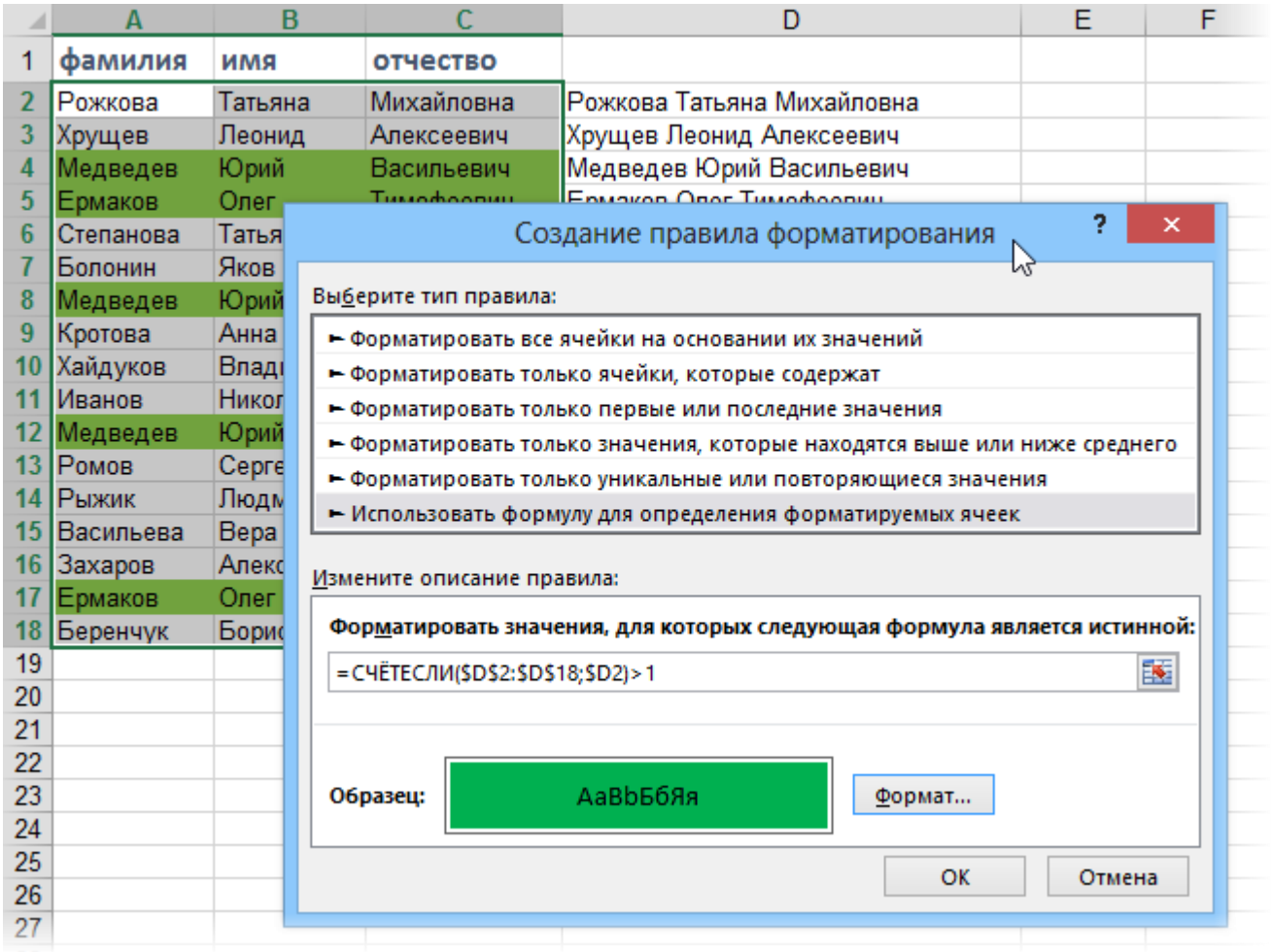
Самым простым решением будет добавить дополнительный служебный столбец (его потом можно скрыть) с текстовой функцией **СЦЕПИТЬ (CONCATENATE)**, чтобы собрать ФИО в одну ячейку:



	A	B	C	D	E
1	**фамилия**	**имя**	**отчество**		
2	Рожкова	Татьяна	Михайловна	Рожкова Татьяна Михайловна	
3	Хрущев	Леонид	Алексеевич	Хрущев Леонид Алексеевич	
4	Медведев	Юрий	Васильевич	Медведев Юрий Васильевич	
5	Ермаков	Олег	Тимофеевич	Ермаков Олег Тимофеевич	
6	Степанова	Татьяна	Алексеевна	Степанова Татьяна Алексеевна	
7	Болонин	Яков	Сергеевич	Болонин Яков Сергеевич	

Имея такой столбец, мы фактически сводим задачу к предыдущему способу.

Как дополнительный вариант, для подсветки совпадающих ФИО можно выделить все три столбца с именами и создать новое правило форматирования, т.е. нажать на вкладке **Главная (Home)** кнопку **Условное форматирование – Создать правило (Conditional Formatting – New Rule)** и выбрать тип правила **Использовать формулу для определения форматируемых ячеек (Use a formula to determine which cells to format)**. Затем ввести формулу проверки количества совпадений и задать цвет с помощью кнопки **Формат (Format)**:



Функция **СЧЁТЕСЛИ (COUNTIF)** в приведенной формуле подсчитывает число вхождений текущего ФИО в диапазон D2:D18. Если это число больше 1, то перед нами дубликат, и мы его заливаем цветом.