Лекция 4. Анализ данных.

# На этой лекции мы:

1. Научимся искать данные в таблицах, фильтровать их, сортировать и удалять дубликаты;
2. Узнаем, что такое сводные таблицы, как их создавать в MS Excel и каким образом можно построить сводные диаграммы;
3. Разберемся, что такое консолидация данных, и как ее можно организовать;
4. Научимся добавлять промежуточные итоги в таблицы с данными.

# Материалы к лекции:

Пример 1.

Пример 2.

Пример 3\_1.

Пример 3\_2.

Презентация.

# На прошлом уроке:

Мы узнали, какие есть инструменты визуализации данных: условное форматирование, диаграммы и спарклайны. Разобрались, как правильно их применять. Научились строить линии тренда на диаграммах и прогнозировать параметры с помощью различных инструментов MS Excel.

# Основные термины:

Поиск по маске – поиск данных с применением маски - комбинации, которая содержит специальные символы, которые заменяют собой неизвестные данные:

\* (звездочка) – заменяет собой любое количество символов (даже если их и вовсе нет, то есть пустую строку)

? (вопросительный знак) – заменяет ровно один любой символ (ни больше ни меньше – один!).

Фильтрация данных – процесс выделения (или отображения) только тех данных, которые удовлетворяют, какому-либо условию в данном случае фильтр – набор значений, которые не должны появляться в конечной выборке.

Сортировка данных – это расположение данных в каком-либо заданном порядке по убыванию или возрастанию параметра (или набора параметров)

Дубликаты – это повторяющиеся строки, повторяющиеся данные, которые, фактически, указывают на совпадающие объекты, события или явления.

Сводная таблица – таблица, в которой в обобщенном виде хранятся данные из одной или нескольких таблиц, как правило в виде отчетности, суммы или среза данных.

Сводная диаграмма – диаграмма, построенная на базе сводной таблицы.

Консолидация данных – объединение данных из разных источников в единую структуру для обобщения и анализа.

Промежуточные итоги – контрольные значения по группам данных.

# План лекции

[Введение](#_Toc106587096)

[Поиск, Сортировка, фильтрация данных и удаление дубликатов.](#_Toc106587097)

[Сводные таблицы и диаграммы.](#_Toc106587098)

[Консолидация данных и создание промежуточных итогов](#_Toc106587099)

[Выводы](#_Toc106587100)

### Введение

Здравствуйте! Приветствую Вас на еще одном, четвертом уроке курса «Электронные таблицы в аналитике»! Мои поздравления, Вы успешно освоили основы работы с электронными таблицами, и мы переходим к теме анализа данных с помощью них! В рамках этого курса мы рассмотрим инструменты MS Excel, не вдаваясь в тонкости процесса непосредственно анализа данных - Вы сможете более подробно изучить этот вопрос позднее во время обучения в GeekBrains или уточнить какие-то моменты, воспользовавшись [списком рекомендованной литературы](#_Рекомендуемая_литература:).

На прошлом занятии мы узнали, какие есть инструменты визуализации данных, а именно: условное форматирование, диаграммы и спарклайны. Также мы научились строить линии тренда на диаграммах и прогнозировать параметры с помощью различных инструментов MS Excel.

Темой урока сегодня являются методы подготовки данных к анализу и некоторые методы их анализа:

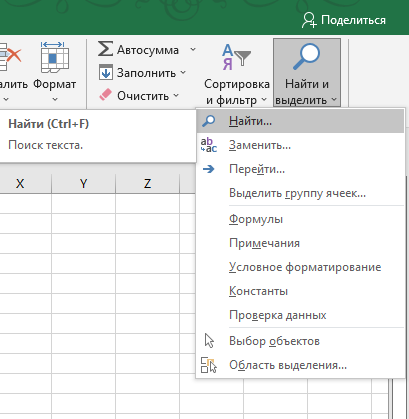
* Подготовка данных – поиск, сортировка, фильтрация и удаление дубликатов;
* Построение сводных таблиц и диаграмм;
* Консолидация данных;
* Создание промежуточных итогов.

### Поиск, Сортировка, фильтрация данных и удаление дубликатов.

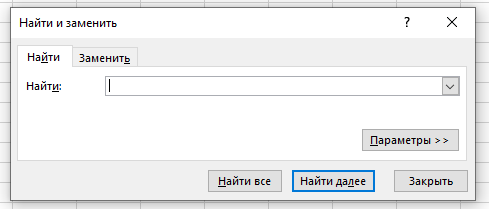
Вы можете выбрать любую таблицу с большим объемом данных, чтобы попробовать функции MS Excel, которые мы рассмотрим далее. Например, таблицу из примера 1 из материалов к лекции.

Поиск данных

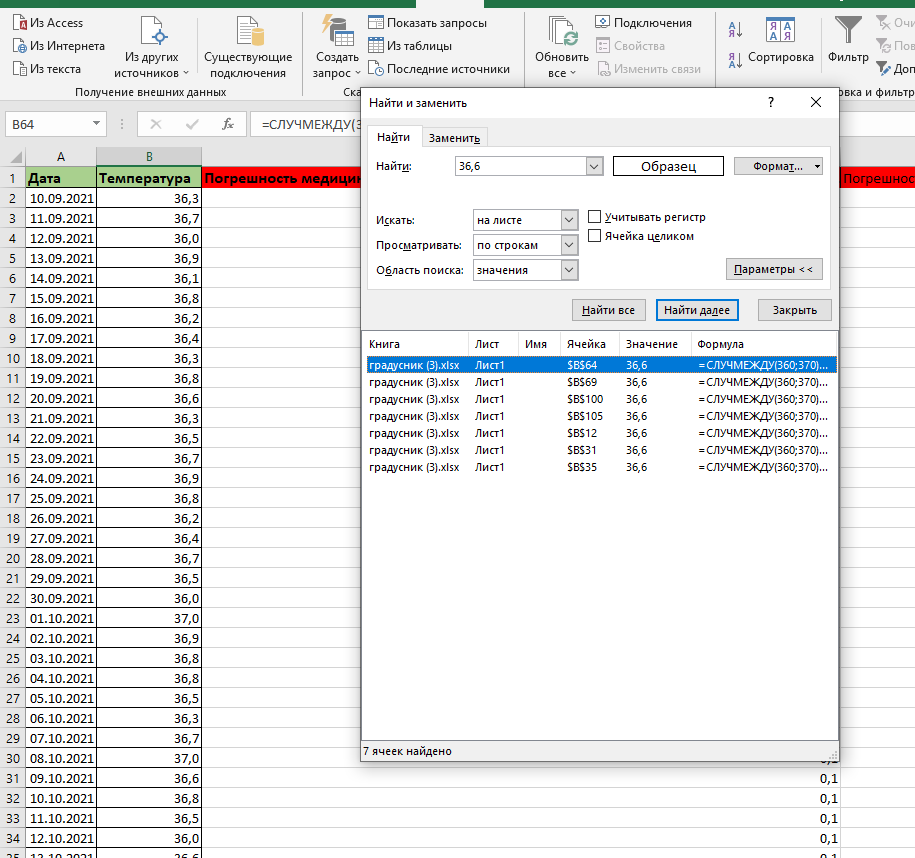
В больших таблицах наиболее ощутимой становится потребность быстро найти нужные данные. Например, Вам нужно найти все заказы, которые были сделаны у заданного поставщика, или выделить объем продаж в конкретном регионе. Самый простой способ это сделать – использовать встроенный инструмент поиска Excel. Его можно найти на вкладке «Главная» >> «Найти и выделить» >> «Найти» или использовать сочетание клавиш **Ctrl+F.**



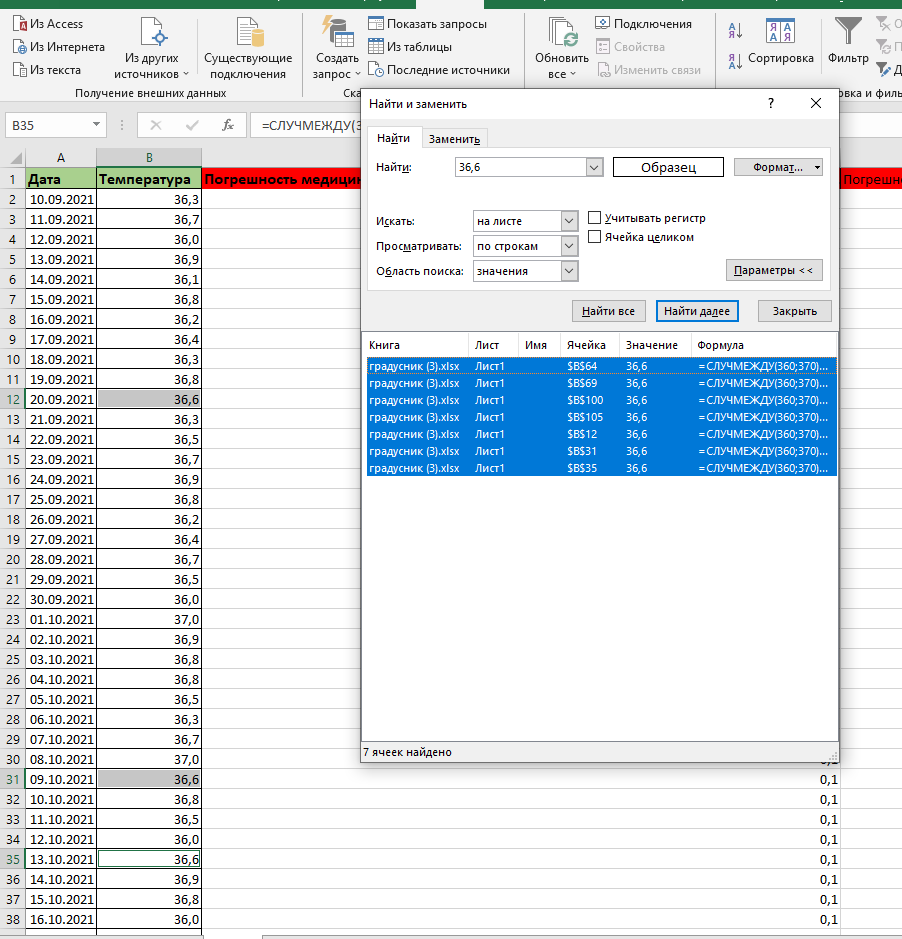
В появившемся окне можно ввести текст, который требуется найти в таблице.



Каждое следующее вхождение искомого текста в таблице можно переключить с помощью кнопки «Найти далее» или получить сразу список всех найденных ячеек с помощью кнопки «Найти все».



Если в списке найденного щелкнуть по любой строке, то произойдет переход к выделенной ячейке. А если выделить все результаты сочетанием клавиш **Ctrl+A**, то выделятся сразу все ячейки с найденными значениями.



Бывает ситуация, когда полная информация неизвестна или у искомых объектов имеется какая-либо общая часть, а другие могут быть разными. Например, нужно найти данные о транзакции определенного типа в какой-то месяц, но точная дата неизвестна. Или нужно выделить все данные с названиями вроде: «Замеры1», «Замеры2», Замеры3»… и так далее.

В таком случае для поиска целесообразно применить маски – это комбинации, которые содержат специальные символы, которые заменяют собой неизвестные:

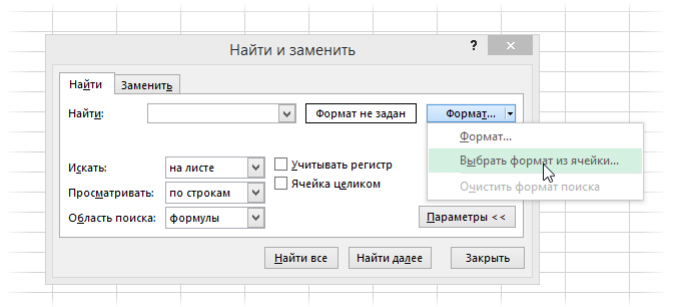
\* (звездочка) – заменяет собой любое количество символов (даже если их и вовсе нет, то есть пустую строку)

? (вопросительный знак) – заменяет ровно один любой символ (ни больше ни меньше – один!)

Примеры использования масок:

|  |  |
| --- | --- |
| \*77 | поиск всех значений, которые заканчиваются на 77 (или состоят только из символов 77); |
| \*aaa\* | поиск всех значений, содержащих три символа «а» подряд в любой их части. |
| doc? | поиск всех значений, которые содержат один символ после doc, например doc1, doc3, docs и т.д. |
| ?? | поиск всех значений, содержащих ровно два символа. |

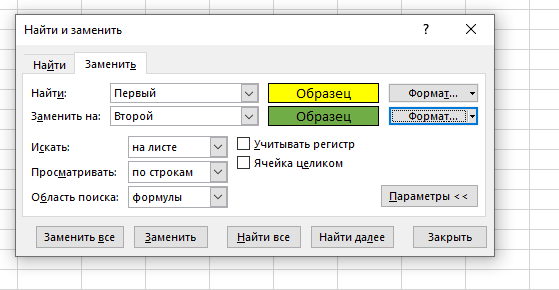
Если нужно произвести поиск еще и с учетом формата ячейки, (например, найти желтые ячейки) можно воспользоваться раскрывающимся списком «Формат»



Он позволит или явно задать формат для поиска или указать ячейку, которую нужно взять за образец.

Например: все ячейки, в которых хранятся данные о сотрудниках отдела «А» выделяются синим цветом. Также, если сотрудник работает в компании меньше года, то его данные выделяются курсивом. И вдруг обнаруживается, что задача должна быть поручена сотруднику отдела «А» и она достаточно простая и наглядная, чтобы с ней справился новичок. Тогда можно найти все ячейки, которые отформатированы синим цветом и содержат данные, написанные курсивом. Полученный набор позволит наглядно посмотреть данные о таких сотрудниках и подобрать исполнителя для задачи.

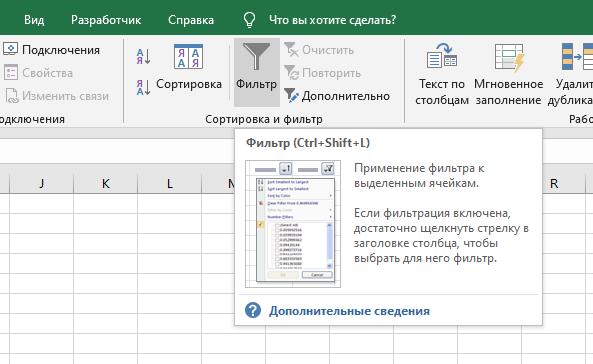
Соседняя вкладка «Заменить» в этом же диалоговом меню поиска позволит не только найти значения, но и сразу выборочно заменить их на заданные. Например, ниже приведена форма, с помощью которой можно заменить все желтые ячейки и слово «Первый» в них на зеленые со словом «Второй».



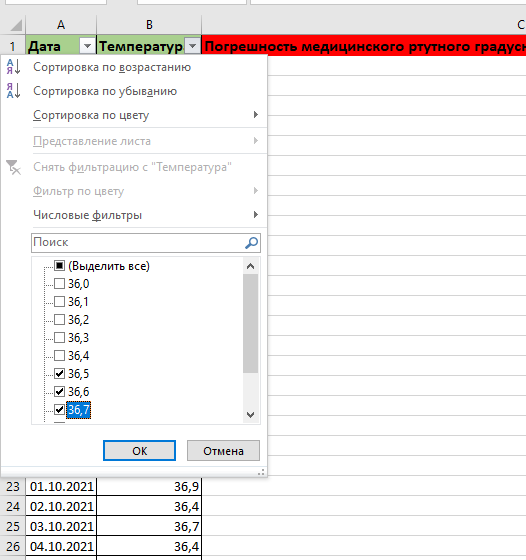
Фильтрация и сортировка данных

Фильтрация данных – процесс выделения (или отображения) только тех данных, которые удовлетворяют, какому-либо условию в данном случае фильтр – набор значений, которые не должны появляться в конечной выборке.

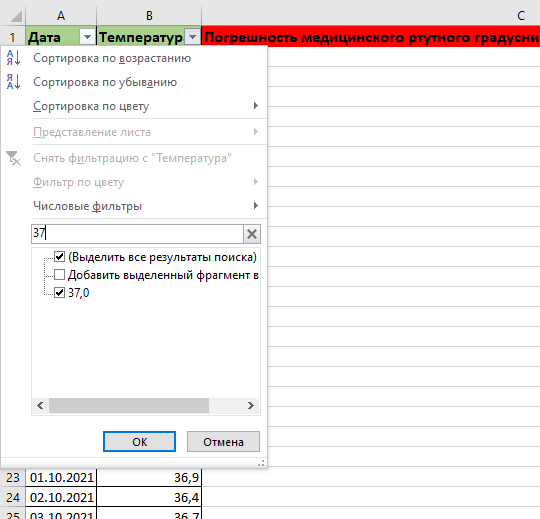
Автофильтр - это один из самых простых в использовании, но одновременно мощных инструментов Microsoft Excel для отбора и последующего анализа данных. Если ваша таблица с данными имеет простую однострочную "шапку", то достаточно просто установить активную ячейку куда-нибудь в данные и перейти на вкладку «Данные» >> «Фильтр».



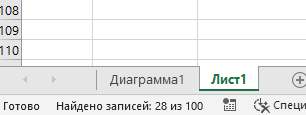
В самом простом случае для фильтрации достаточно щелкнуть по черному треугольнику раскрывающегося списка фильтра в нужном столбце и поставить флажки (галочки) напротив тех элементов, которые мы хотим отобразить:



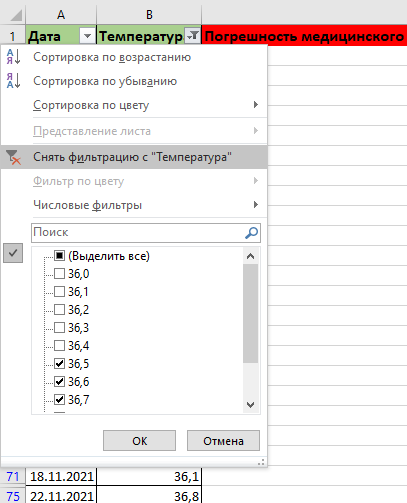
В случае большого списка, чтобы не возиться с галочками, можно ввести несколько символов в поле «Поиск», и Excel отфильтрует те элементы, где содержится введенный текст как подстрока:



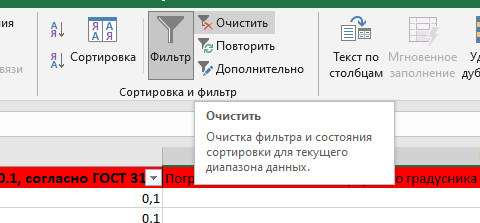
Флажок «Добавить выделенный фрагмент в фильтр» отсеивает данные, содержащие заданный фрагмент, т.е. оставит видимыми только те элементы, где не содержится введенная в поле поиска фраза. Сразу после фильтрации в левом нижнем углу экрана всегда можно найти количество оставшихся после фильтрации строк («Найдено записей..»):



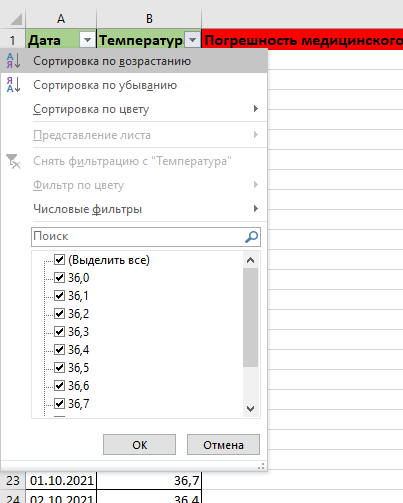
Для сброса отдельного фильтра можно выбрать команду «Удалить фильтр» со столбца:



Для сброса сразу всех фильтров можно воспользоваться кнопкой «Очистить» на вкладке «Данные».



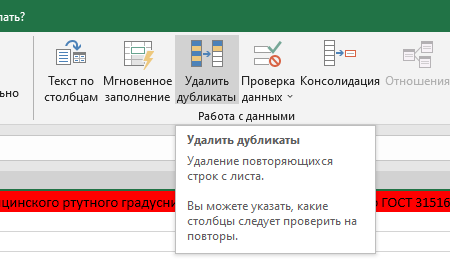
В том же контекстном меню можно воспользоваться сортировкой строк по возрастанию и убыванию значений в выбранном поле:



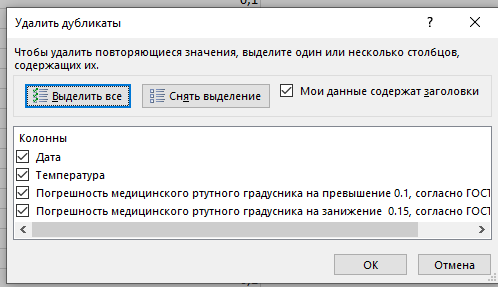
Сортировка данных – это расположение данных в каком-либо заданном порядке по убыванию или возрастанию параметра (или набора параметров)

Удаление дубликатов

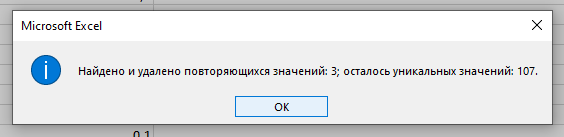
Зачастую данные могут поступать одновременно из разных источников, и в этом случае высока вероятность появления дублирующихся строк. Дубликаты – это повторяющиеся строки, повторяющиеся данные, которые, фактически, указывают на совпадающие объекты, события или явления. В такой ситуации постоянно приходится заниматься удалением появившихся дубликатов. В MS Excel это можно осуществить с помощью инструмента «Удалить дубликаты» во вкладке «Данные»:



Выделите любую ячейку в обрабатываемом диапазоне данных. Нажмите на команду «Удалить дубликаты» и откроется диалоговое окно:



Здесь можно флажками отметить столбцы, значения из которых будут учтены при поиске повторяющихся значений. По умолчанию выбраны все столбцы. Если кликнуть на кнопку «ОК», то Excel просмотрит все строки с повторяющимися значениями и выведет сообщение о том, сколько дубликатов было удалено: при этом одно из дублирующихся значений останется в таблице, как уникальное.



Если вы совершили ошибку при удалении дубликатов, Вы всегда можете отменить это действие с помощью комбинации клавиш **Ctrl+Z**.

### Сводные таблицы и диаграммы.

Во многих ситуациях в сфере бизнеса требуется проанализировать данные или создать срез данных, для того чтобы сделать важные выводы. Например, представьте, что вы продаете продовольственные товары в разных магазинах в разное время. Вам нужно отслеживать тысячи точек данных. Обобщить такие данные так быстро, насколько это вообще возможно, позволяют сводные таблицы. Сводная таблица – таблица, в которой в обобщенном виде хранятся данные из одной или нескольких таблиц, как правило в виде отчетности, суммы или среза данных.

Примеры:

*Для данных по продуктовым магазинам с помощью сводной таблицы можно быстро определить следующее:*

1. *Сумму, оставленную покупателями за год в каждом магазине за каждый продукт;*
2. *Объем выручки для каждого магазина;*
3. *Объем выручки для каждого года.*

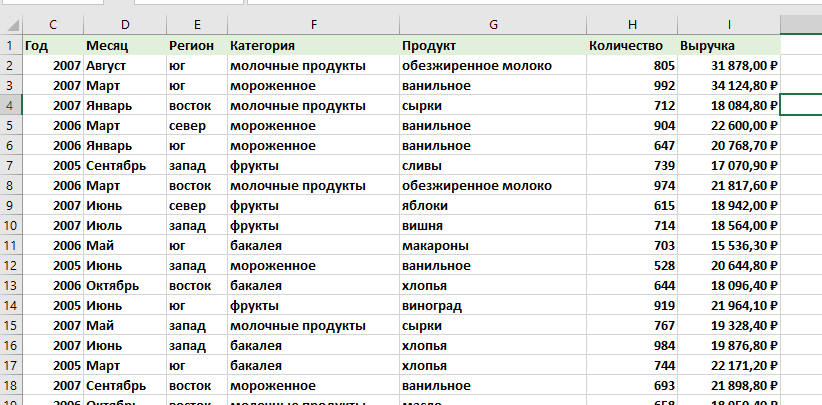
*Для турфирмы можно создать срез данных, позволяющий определить, зависит ли средняя сумма, затрачиваемая на путешествия от возраста, пола или обоих факторов.*

*При анализе продаж автомобилей можно сравнить долю многодетных семей, покупающих автомобиль-универсал с долей небольших семей, приобретающих такой автомобиль.*

Сводная таблица является очень эффективным инструментом для получения среза данных.

Откроем файл примера 2 из материалов к лекции.

Вы увидите таблицу, которая содержит более 900 строк записей. В этих записях указан год, месяц, регион, где расположен магазин. Затем идет категория продуктов, вид продукта, количество проданных единиц и суммарная выручка за указанный период в данном регионе.



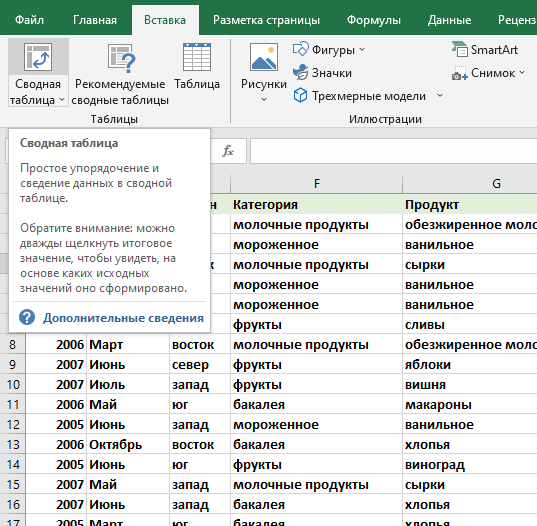
**Задачи:**

1. Построить сводную таблицу, которая содержит данные о продажах молочных продуктов весной 2007 года.
2. Построить сводную диаграмму, отражающую продажи молочных продуктов с 2005 по 2007 год.

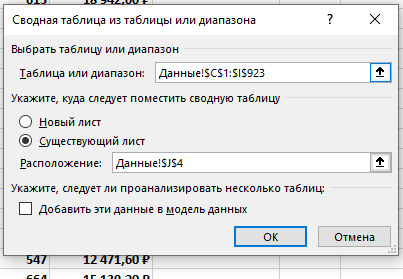
**Задача 1:**

*Построить сводную таблицу, которая содержит данные о продажах молочных продуктов весной 2007 года.*

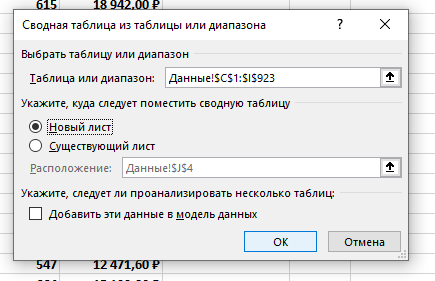
Начнем с построения сводной таблицы. Для этого на вкладке «Вставка» кликнем «Сводная таблица»:



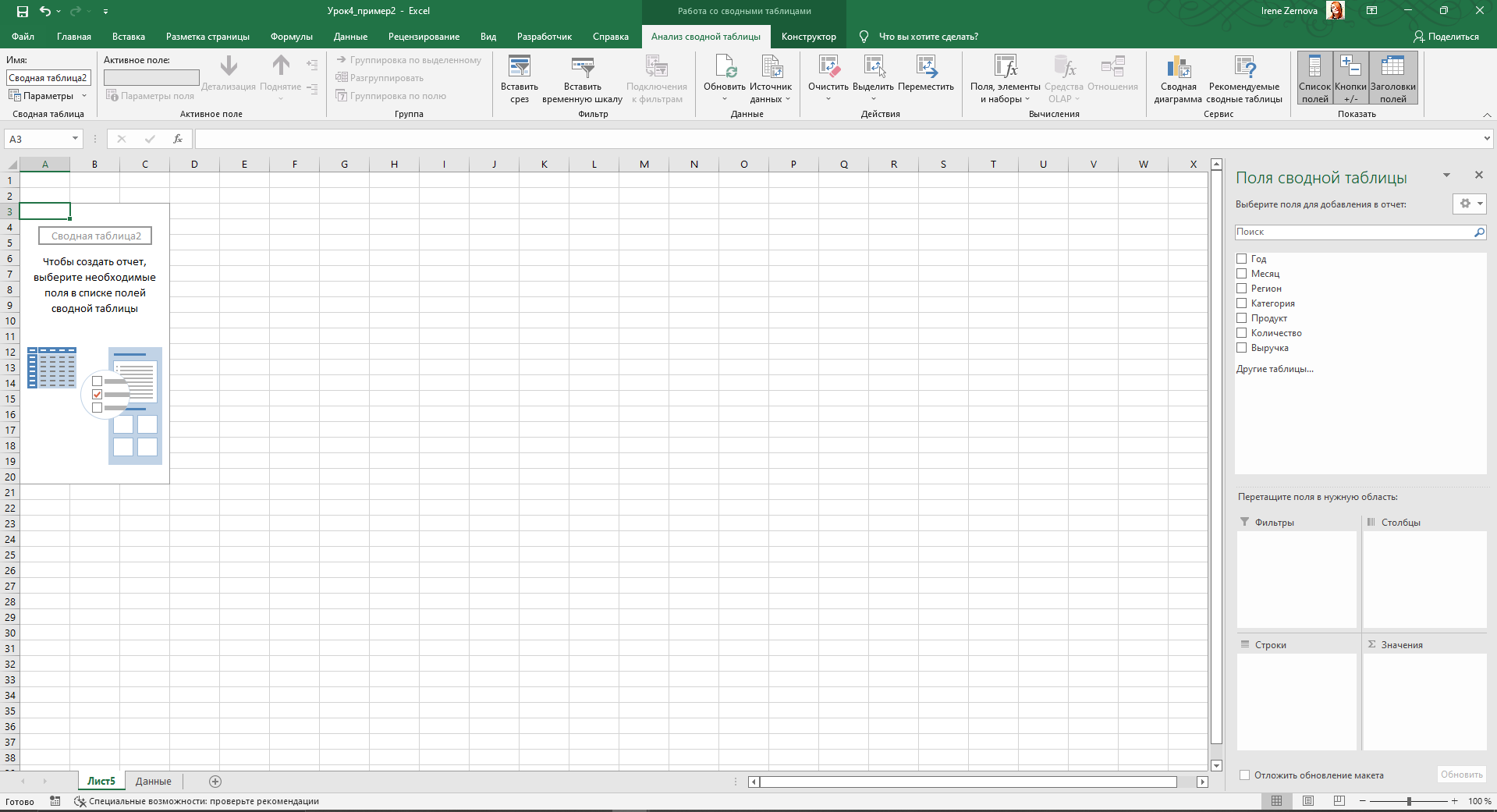
В появившемся диалоговом окне Excel может предложить диапазон данных самостоятельно или его можно указать вручную:



Укажем, что таблицу необходимо разместить на новом листе и кликнем «ОК»:



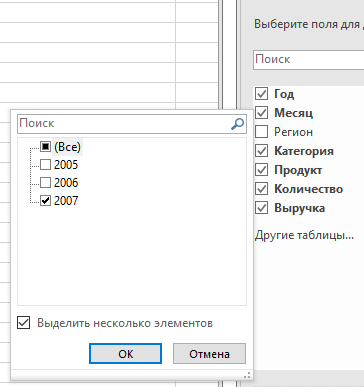
Вы увидите, на новом листе конструктор сводных таблиц в меню справа:



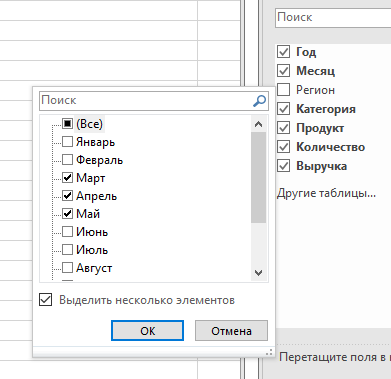
Выберем столбцы, которые должны быть включены в сводную таблицу и отфильтруем параметры в них следующим образом:

* Нам потребуются только те строки, которые имеют значение 2007, в столбце «Год», и значения «Март», «Апрель» или «Май» в столбце «Месяц». Значит зададим два фильтра:

Год: 2007



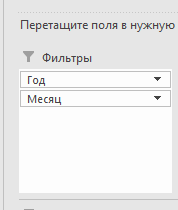
Месяц: Март, Апрель, Май



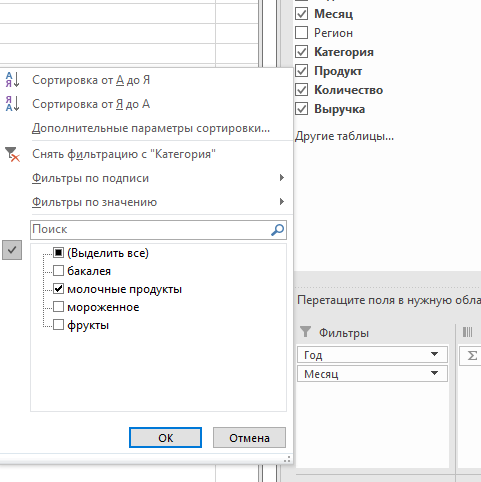
*В режиме конструктора можно создавать таблицу если «перетаскивать» столбцы основной таблицы в поля четырех параметров:*

1. *Фильтры – нужен, чтобы выборочно отображать данные в сводной таблице;*
2. *Строки и столбцы – устанавливает, из какого столбца будут взяты значения для заголовков таблицы. Здесь можно установить иерархию заголовков, перемещая столбцы ниже или выше в самом поле.*
3. *Значения – устанавливает, какие значения требуется отобразить в ячейках таблицы. Иногда дополняется автоматически при указании других параметров.*

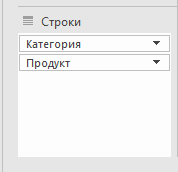
* Добавим столбцы «Год» и «Месяц» в графу «Фильтры» сводной таблицы:

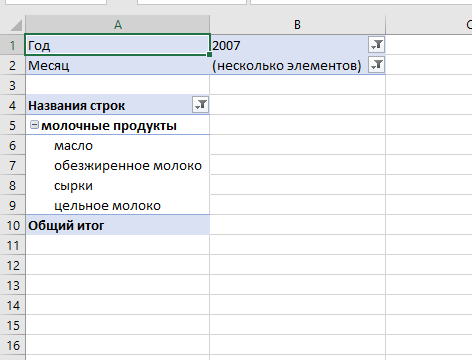


* Строкой сводной таблицы должно быть значение столбца «Категории»: отфильтруем только те, которые являются «молочными продуктами»:

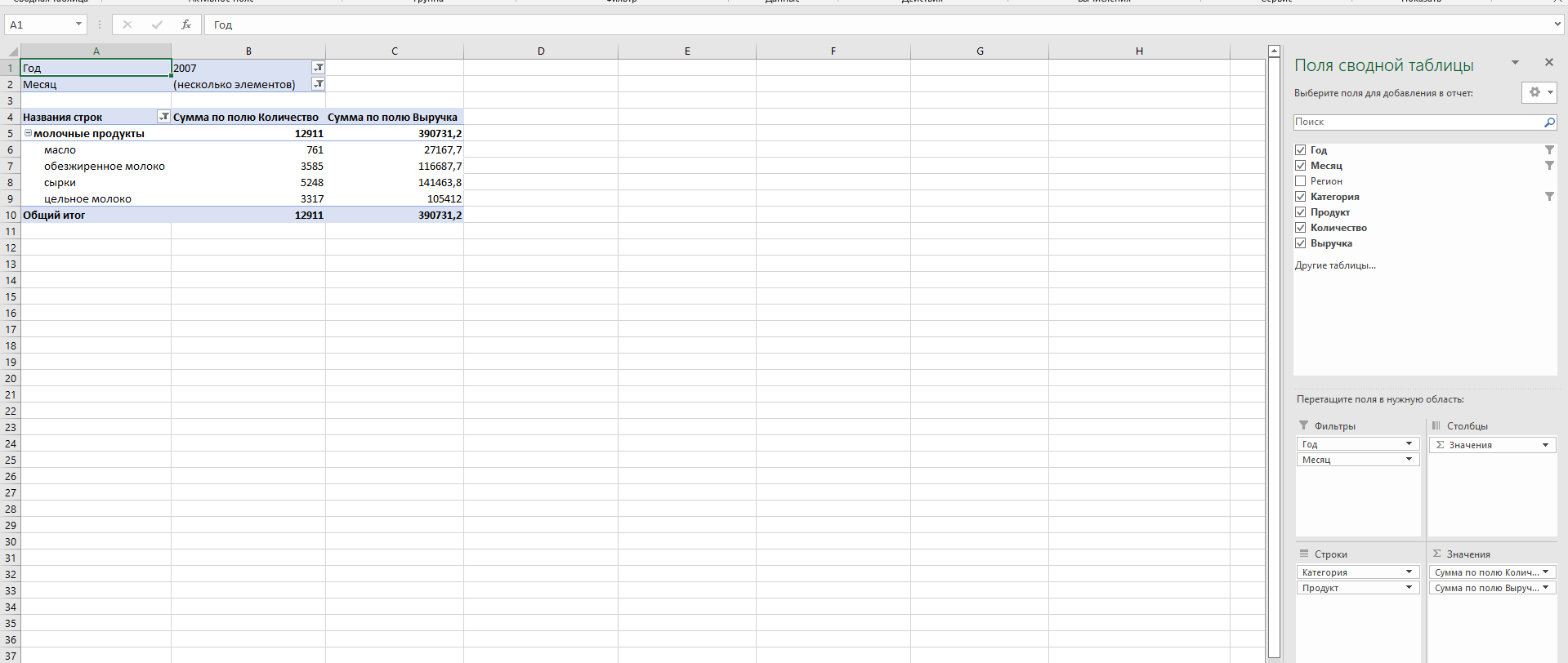


Добавим этот столбец в графу «Строки» сводной таблицы. Под ним можно разместить еще столбец «Продукты». Тогда разные продукты будут пунктами основной строки:





* Теперь укажем, какие значения нужно отобразить в сводной таблице. Перенесем значения из столбцов «Количество» и «Выручка» в графу «Значение» сводной таблицы:



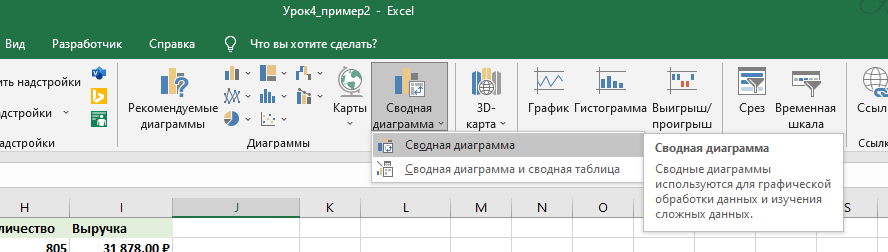
Таким образом нам удалось выделить данные из довольно большого объема и создать емкий отчет, обобщающий их.

**Задача 2:**

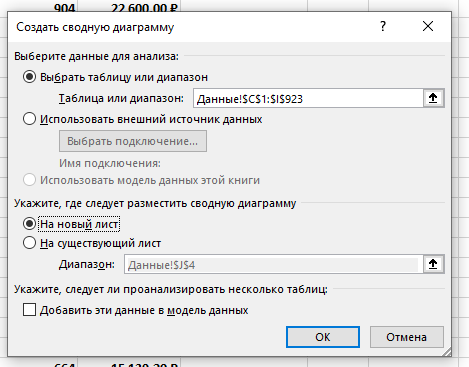
*Построить сводную диаграмму, отражающую продажи молочных продуктов с 2005 по 2007 год.*

Теперь построим сводную диаграмму по выручке за молочные продукты с 2005 по 2007 год.

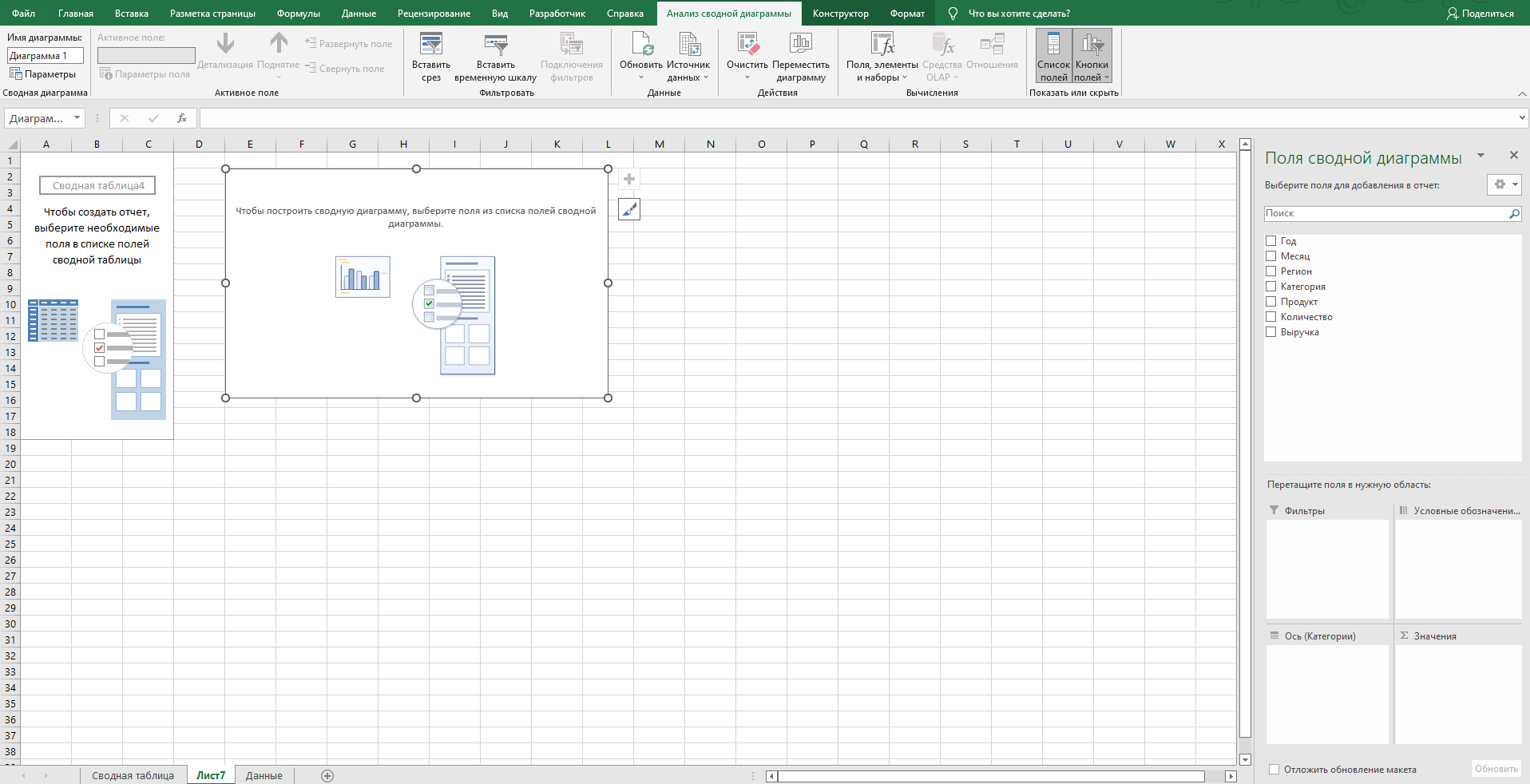
Во вкладке «Вставка» выбираем вставку сводной диаграммы:



Указываем диапазон данных и построение диаграммы на новом листе:

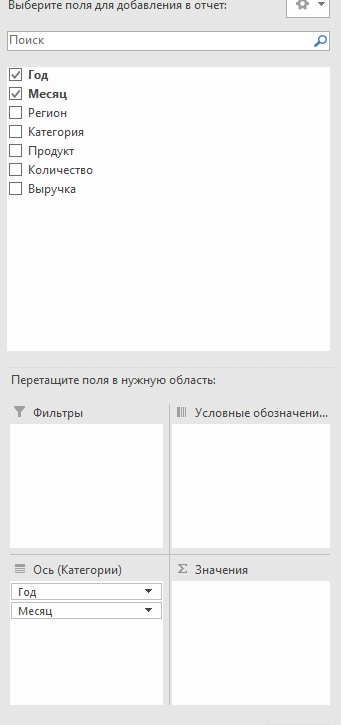


Увидим здесь конструктор диаграмм, очень похожий на конструктор сводных таблиц:

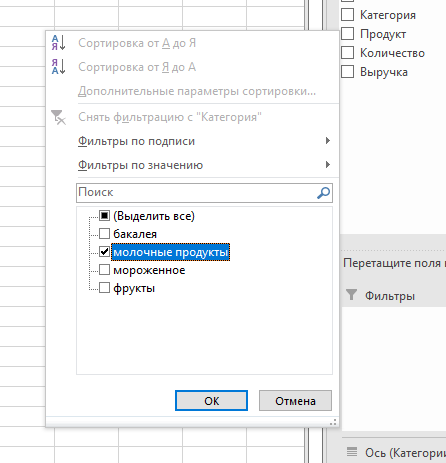


Выбираем параметры следующим образом:

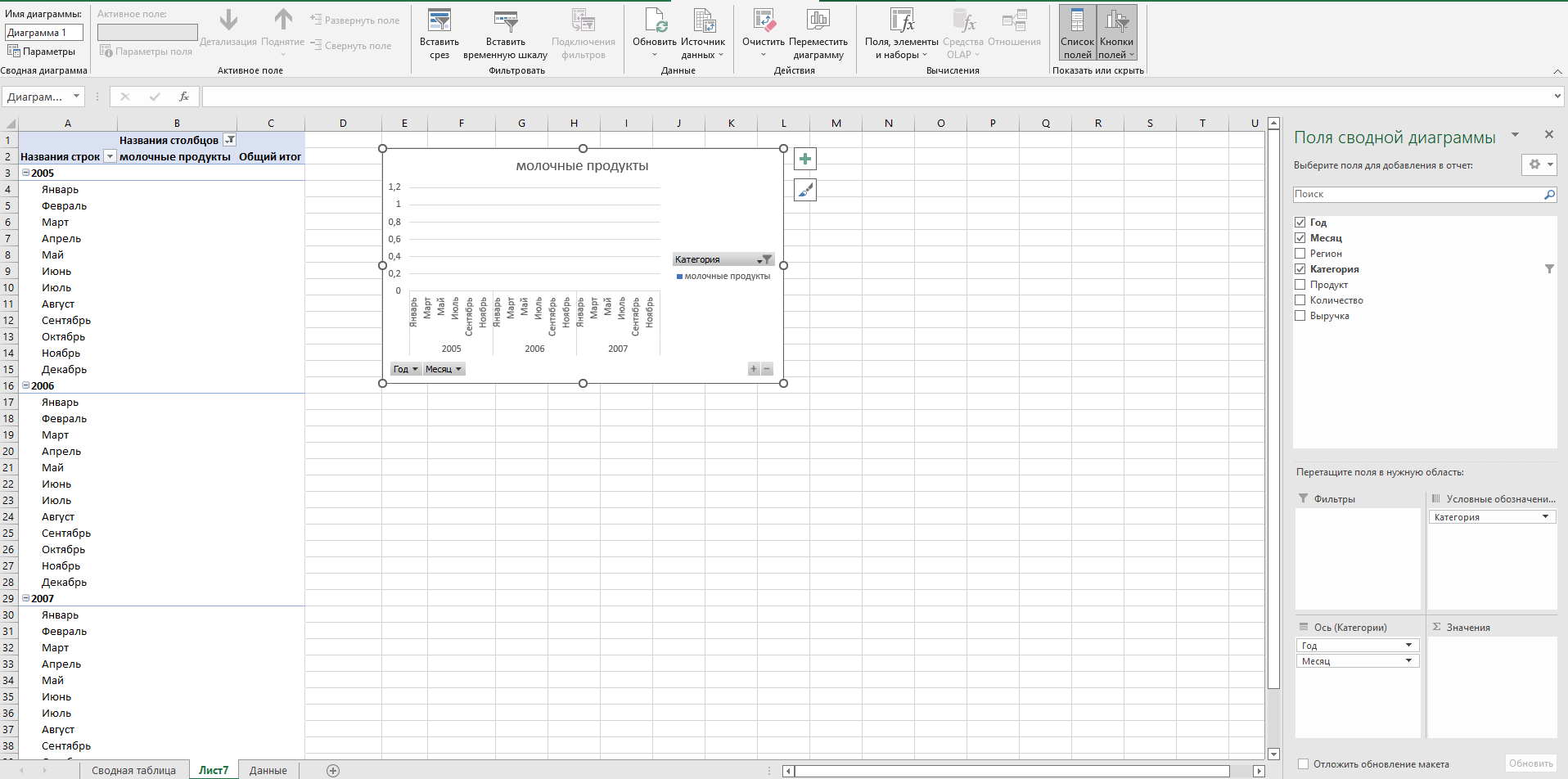
* Нам необходимы данные за 2005-2007 год, поэтому добавим столбцы «Год» и «Месяц» в поле «Ось категорий»



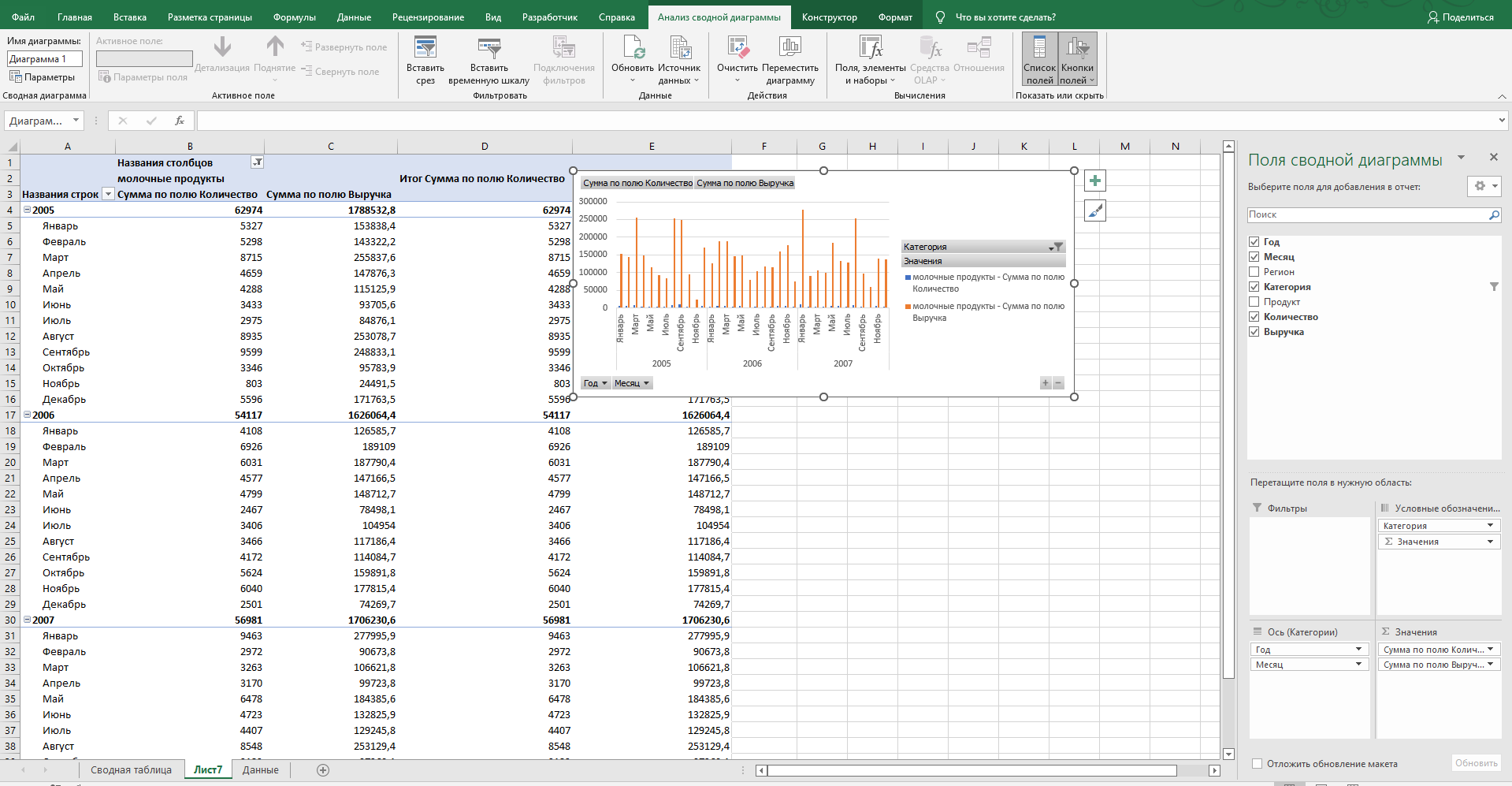
* Категория нам потребуется только «молочные продукты» - отфильтруем категорию:



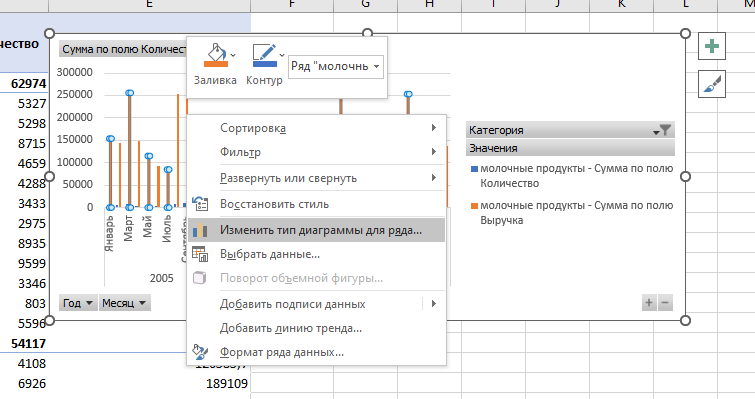
* Перетащим полученный столбец «Категории» в поле «Условные обозначения»:

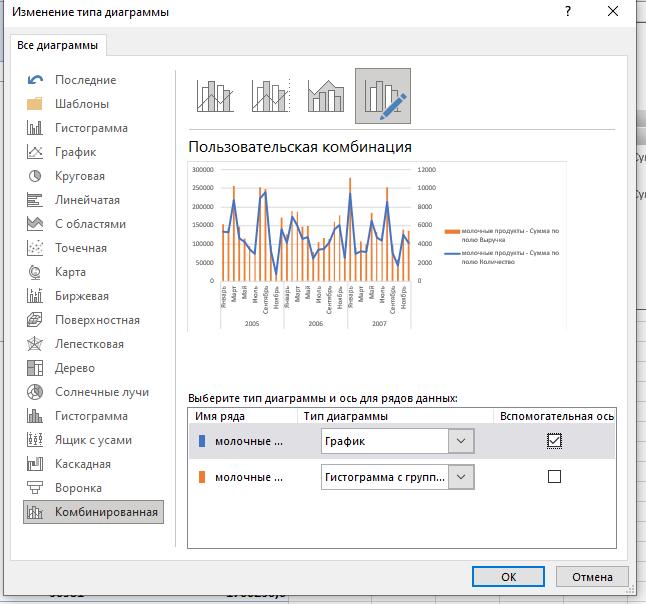


* Затем выберем, какой ряд значений будет отображен на диаграмме – поместим в поле «Значение» столбцы «Количество» и «Выручка»:

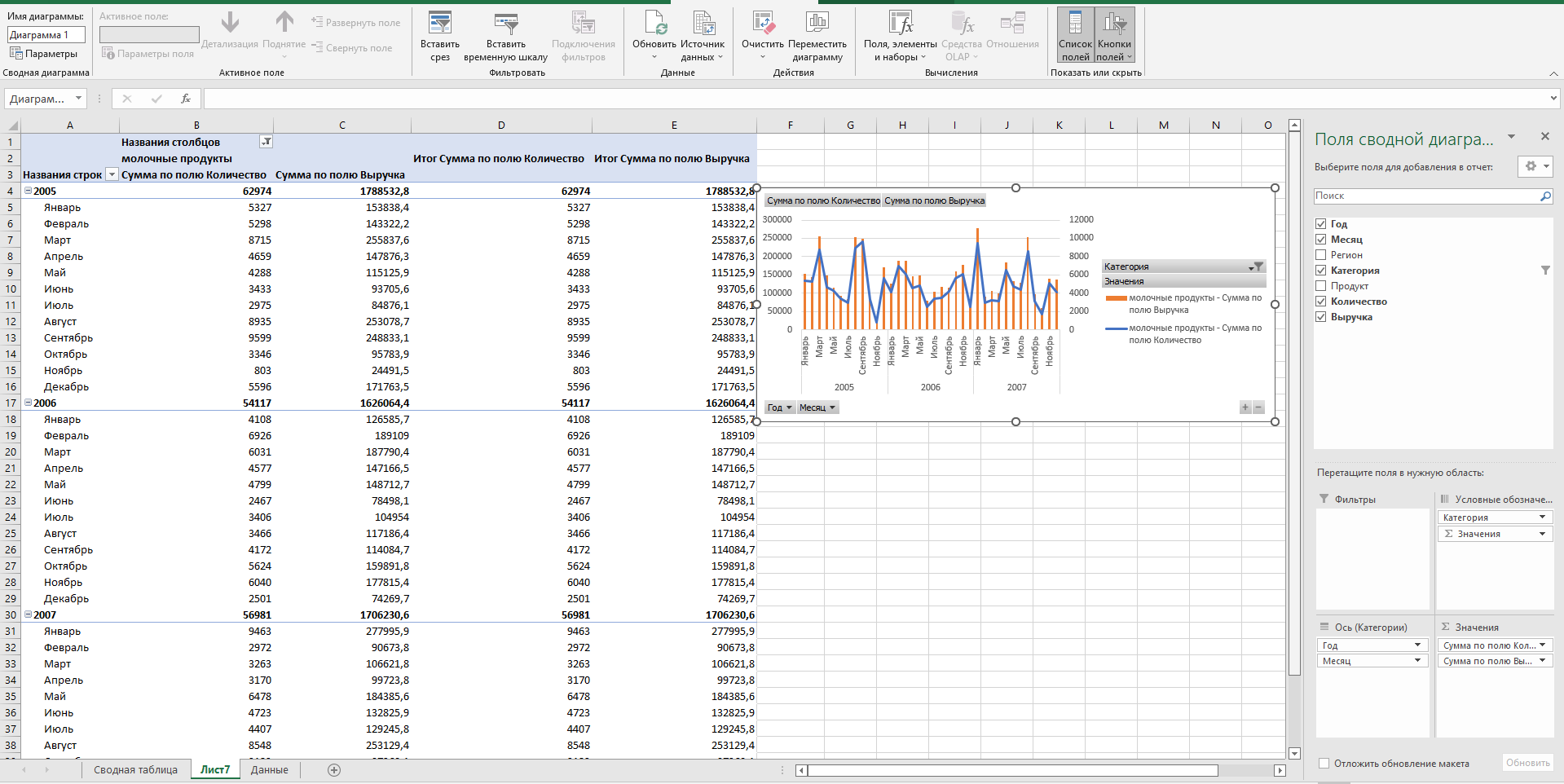


* Изменим тип полученной диаграммы на комбинированную (график + гистограмма – таким образом, с учетом масштаба, величины будут отражены более наглядно и такой тип комбинированной диаграммы позволит оценить зависимость количества проданных единиц и выручки) и добавим вспомогательную ось:





И видим, что построена сводная диаграмма и сводная таблица, данные которой на ней отражены:



Можно сделать вывод, что построение диаграмм на базе выборки значений, а не ряда целиком, заметно упрощается благодаря инструменту построения сводных диаграмм.

Таким образом, сводные диаграммы позволяют легко обобщать данные и получать их срезы для анализа больших объемов записей.

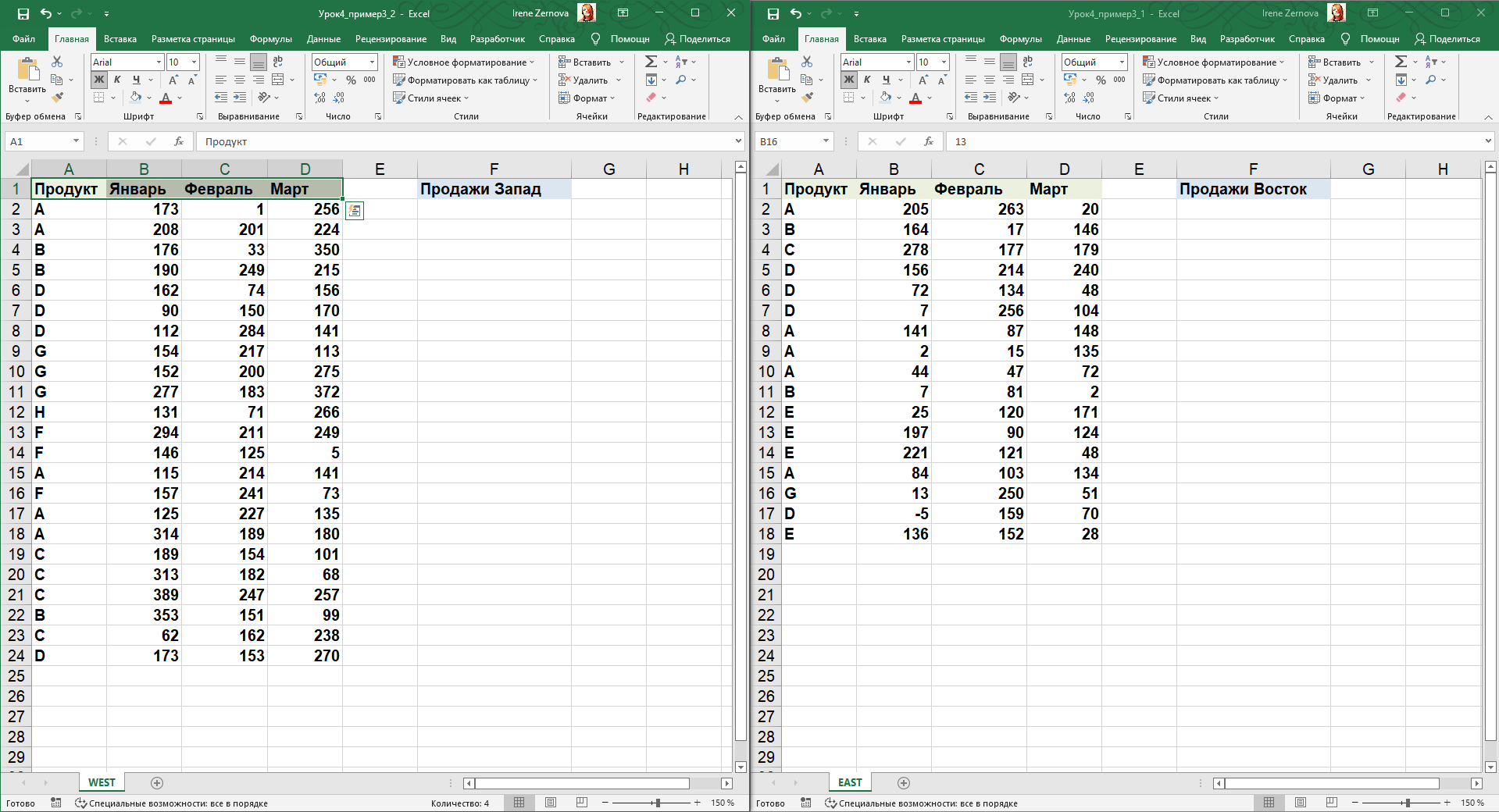
### Консолидация данных и создание промежуточных итогов

Бизнес-аналитики часто получают листы с одними и теми же обобщенными данными (такими как ежемесячные продажи продукции) из различных филиалов или регионов. Аналитику, как правило, приходится объединять или консолидировать эту информацию в одной книге MS Excel для определения общей прибыльности компании. Консолидация данных – объединение данных из разных источников в единую структуру для обобщения и анализа. Для этой цели можно использовать инструмент «Консолидация» на вкладке «Данные». С его помощью можно обеспечить автоматическое отражение на объединенном листе изменений, произошедших на отдельных листах.

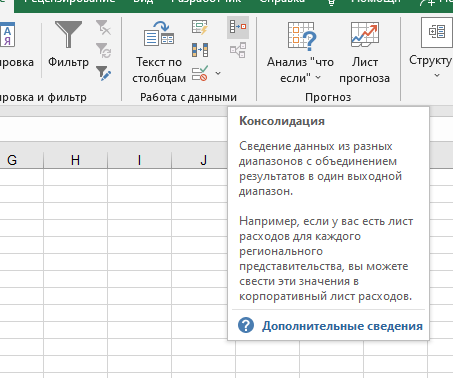
Откройте файлы: Пример3\_1 и Пример3\_2 из материалов к лекции. На них вы увидите данные о ежемесячных продажах продуктов на востоке и на западе страны.

**Задача:** Создать объединенный лист, где будут сведены в таблицу общие объемы продаж для каждого продукта за месяц.

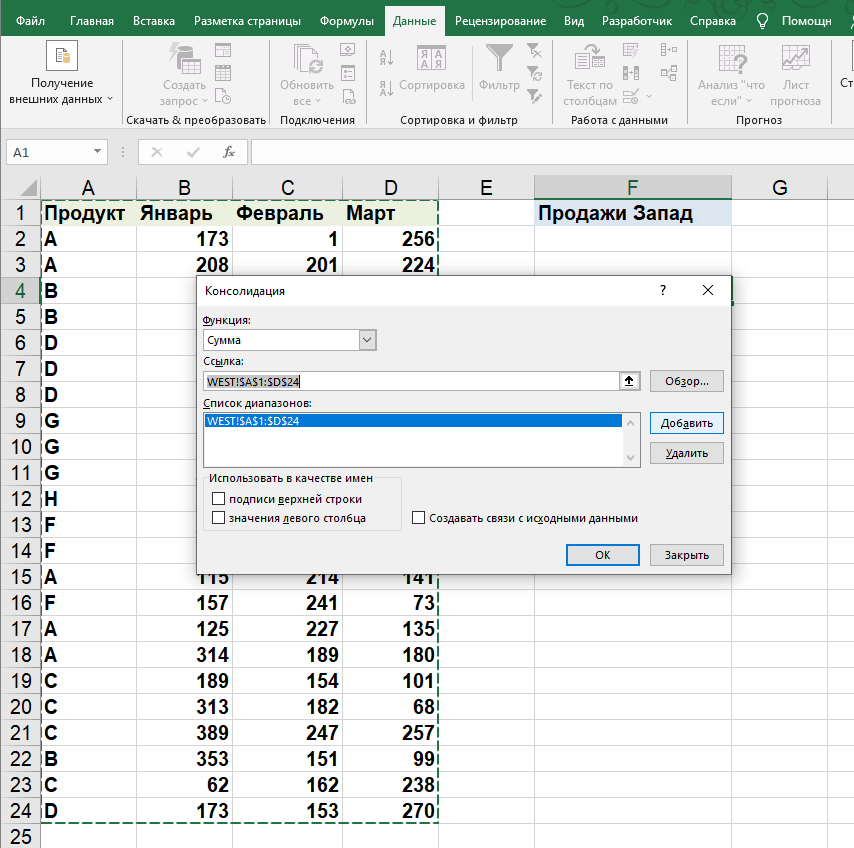
Для начала разместите для удобства оба листа на экране так, чтобы они были видны одновременно (это можно сделать, открыв оба файла, сделайте одно из окон активным и нажмите клавишу **Win+стрелка вправо** (или нажмите на значок  на macOS и удерживайте его до появления контекстного меню). У вас разделится экран и для отображения на второй половине выберете второй лист с данными.



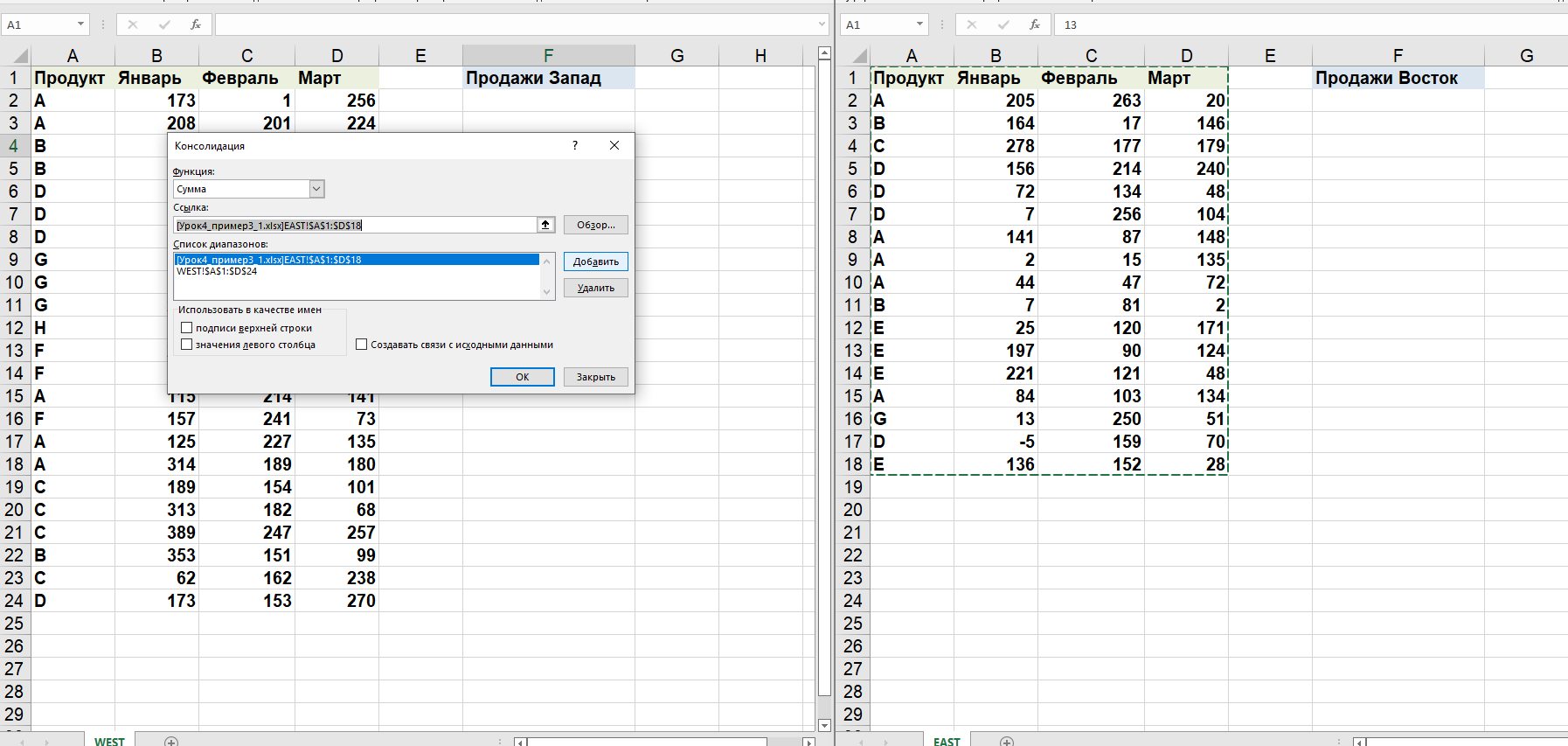
Создайте и откройте пустой лист. Во вкладке «Данные» кликните на значок инструмента «Консолидация»:



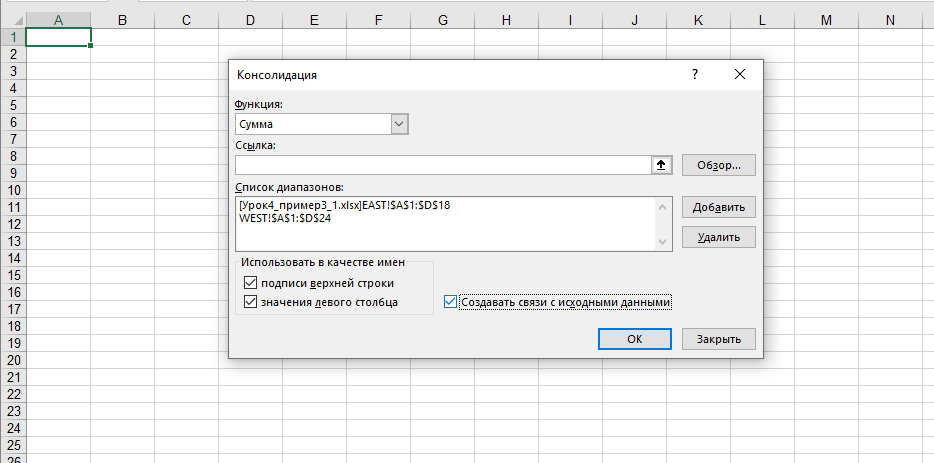
В появившемся диалоговом окне Вам потребуется по очереди выбрать диапазоны для объединения на каждом листе. То есть, выбираем диапазон А1:D24 с листа с продажами на западе, затем нажимаем кнопку «Добавить».



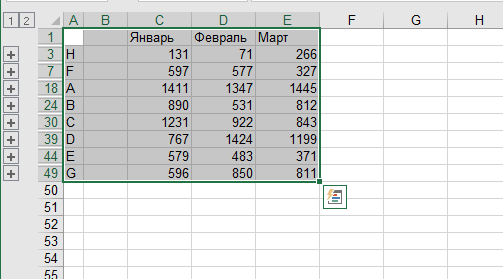
Затем выбираем мышкой диапазон А1:D18 с другого листа – у вас отобразится ссылка, содержащая название файла, который содержит выделенный диапазон – и нажмите кнопку «Добавить»:



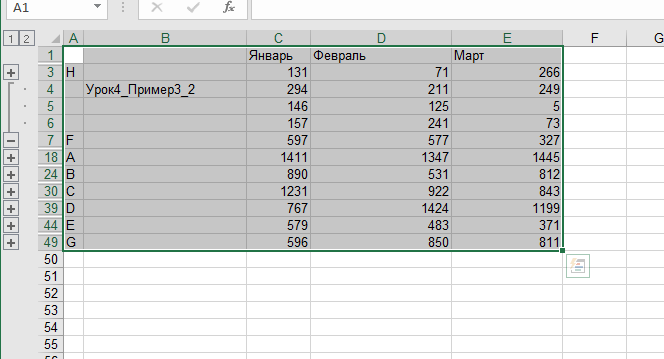
Отметим флажок «Подписи верхней строки» и «значения левого столбца» для имен – это показывает, что данные объединяются путем сопоставления значений (или меток) в верхней строке и левом столбце каждого диапазона. Также можно поставить флажок в параметре «Создавать связи с исходными данными» – так вы обеспечите отображение изменений в выделенных диапазонах на объединенном листе, то есть если внести изменения в один из листов, значения в сводной таблице, которые с ними связаны, тоже изменятся. Заполненная форма для консолидации должна выглядеть следующим образом:



Нажимаем «Ок» и получаем сводную таблицу с объединенными данными:

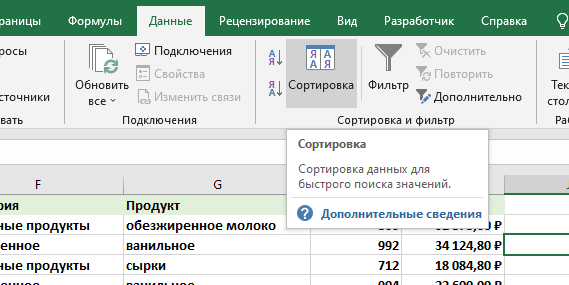


Здесь можно нажимать на плюсики и раскрывать строки, которые «свернуты»:

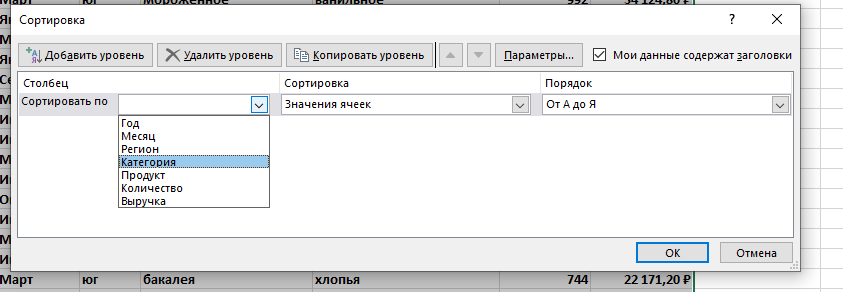


Однако, зачастую требуется получить не срез данных, а просто обработать список и получить какие-то промежуточные итоги. Промежуточные итоги – контрольные значения по группам данных. Например, суммировать выручку по регионам, по продуктам, по продавцам.

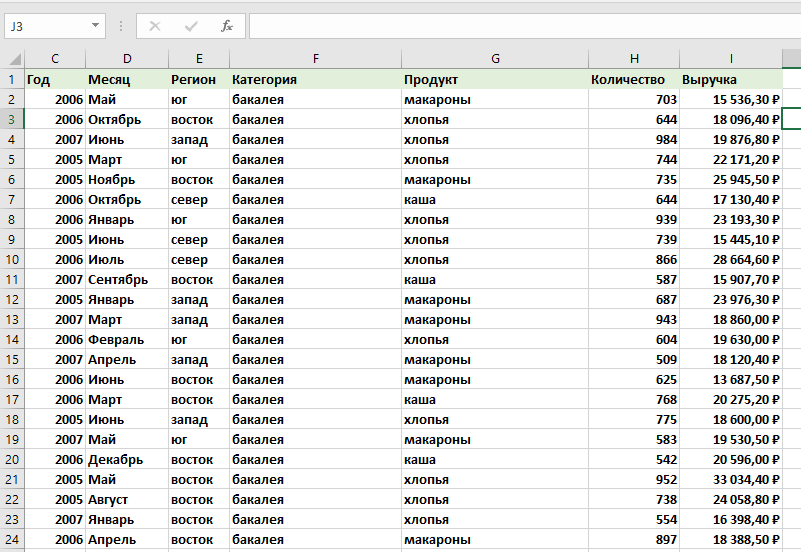
Если отсортировать список по столбцу с нужными данными, то с помощью инструмента «Промежуточный итог» можно создать в списке итоги по группам данных. Например, если базу данных косметических товаров отсортировать по регионам, то можно вычислить общую выручку и количество единиц продукции для каждого региона и поместить промежуточные данные непосредственно под последней строкой для соответствующего региона. Откроем снова файл Примера 2 из материалов к лекции. Отсортируем таблицу по столбцу «Категория». ВО вкладке «Данные» кликнем по инструменту «Сортировка».



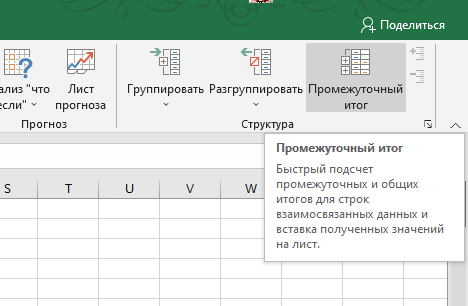
В диалоговом окне выберем сортировать по значению из столбца «Категория»:



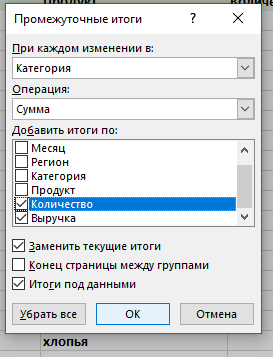
Нажимаем «Ок» и видим таблицу, отсортированную по значению из столбца «Категория»:



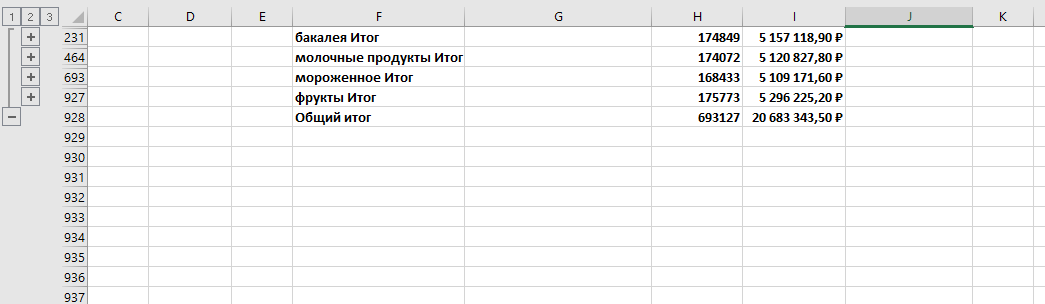
Теперь во вкладке «Данные» выберем инструмент «Промежуточный итог»:

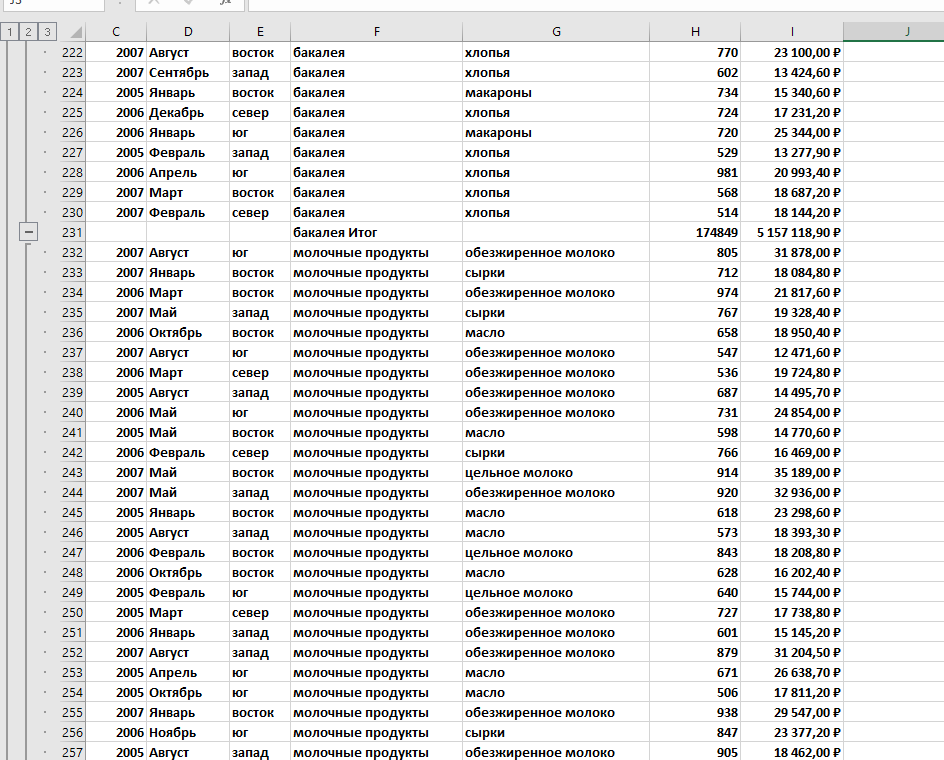


Укажем, что итог нужно добавить при изменении значения в поле «Категория». Операция суммирования вернет сумму по полям, отмеченным флажками. Отметим поля «Выручка» и «Количество»:



Нажимаем «Ок» и видим, что в таблицу добавились промежуточные итоги и навигация сбоку с помощью вкладок:





### Выводы

На этом уроке мы познакомились с инструментами обработки и анализа данных в MS Excel.

Исходя из полученных знаний и опыта можно сделать следующие выводы:

* Excel имеет инструменты, которые значительно сокращают время обработки данных и усилия, которые для этого затрачиваются: без них пришлось бы использовать более сложные комбинации действий с применением формул и вычислений.
* Excel позволяет наглядно создавать сводные таблицы и диаграммы с помощью удобных конструкторов – это освобождает от изучения сложного синтаксиса и позволяет сосредоточится на структуре создаваемых элементов.
* Excel в ходе применения инструментов «Промежуточный итог» и «Консолидация данных» позволяет оптимизировать структуру таблицы, делая ее отображение более корректным – это экономит время и повышает читаемость данных.

# В следующей лекции:

Мы узнаем, что такое надстройки MS Excel как управлять ими. Познакомимся с надстройками, которые входят в стандартный пакет MS Excel и позволяют производить подбор параметра и другие действия, а также узнаем, что такое анализ «что если» и как его использовать.

# Рекомендуемая литература:

1. Винстон У. Бизнес-моделирование и анализ данных: решение актуальных задач с помощью Microsoft Excel: [перевод с английского Ю. Бочина], 5-е изд., Питер, 2019. - 864 с. - (Microsoft).; ISBN 978-5-4461-1339-2
2. [Руководство пользователя Microsoft Excel](https://support.microsoft.com/ru-ru/topic/%D1%81%D0%BF%D1%80%D0%B0%D0%B2%D0%BA%D0%B0-%D0%B8-%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5-%D0%BF%D0%BE-excel-55d6b75e-7f27-4fe6-929c-a6db896dfab6)
3. Павлов Н. Excel – готовые решения. Бери и пользуйся!, 2014. – 382 с., ISBN: 978-5-519-01837-1
4. М. Александер, Р. Куслейка, Д. Уокенбах Excel 2019. Библия пользователя», «ДИАЛЕКТИКА», 2019. - 1136 с.; ISBN 978-5-907144-44-6,

# Использованная литература:

1. [Руководство пользователя Microsoft Excel](https://support.microsoft.com/ru-ru/topic/%D1%81%D0%BF%D1%80%D0%B0%D0%B2%D0%BA%D0%B0-%D0%B8-%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5-%D0%BF%D0%BE-excel-55d6b75e-7f27-4fe6-929c-a6db896dfab6)
2. Винстон У. Бизнес-моделирование и анализ данных: решение актуальных задач с помощью Microsoft Excel: [перевод с английского Ю. Бочина], 5-е изд., Питер, 2019. - 864 с. - (Microsoft).; ISBN 978-5-4461-1339-2
3. М. Александер, Р. Куслейка, Д. Уокенбах Excel 2019. Библия пользователя», «ДИАЛЕКТИКА», 2019. - 1136 с.; ISBN 978-5-907144-44-6,