АОУ ВО ДПО «Вологодский институт развития образования»

Центр непрерывного повышения профессионального мастерства педагогических работников в г. Вологда

«ОДОБРЕНО»

на заседании экспертной рабочей группы

по учебному предмету «Информатика»

при РУМО по общему образованию

(Протокол №4 от 23.06.2025 г.)

**Методический кейс**

**«Методические подходы к формированию у обучающихся умений решения задач по теме «Анализ данных»**

*Автор составитель*

*Ганичева Елена Михайловна,*

*методист сектора естественно-научного*

*и технологического образования ЦНППМПР*

*в г. Вологде АОУ ВО ДПО «ВИРО»*

2025 год

**Аннотация**

В рекомендациях представлен методический кейс учебных материалов для формирования у обучающихся представления о понятии «анализ данных» корреляционном анализе данных, линейном регрессионном анализе данных и задаче кластеризации; формирования умений вычислять коэффициенты корреляции между переменными, устанавливать статистическую взаимосвязь между признаками, строить математическую модель линейной зависимости и с её помощью предсказывать ожидаемое среднее значение переменной, решать задачу кластеризации данных. Кейс предполагает достижение обучающимися предметных результатов: понимания основных принципов анализа данных, умения выполнять последовательность решения задач анализа данных в электронной таблице.

Представлена модель задания №27 из ЕГЭ по информатике, рассмотрены подходы к решению данного задания, приведены решения заданий и подборка таких задач разного уровня сложности. Методический кейс адресован учителям информатики, работающим по программам среднего общего образования.

**Содержание**

1. **Актуальность**

Основной целью изучения учебного предмета «Информатика» на уровне среднего общего образования является обеспечение развития информационных компетенций обучающегося, его готовности к жизни в условиях развивающегося информационного общества. В рамках курса информатики учащиеся имеют возможность получить представление о сквозных цифровых технологиях, таких, как дополненная и виртуальная реальность, робототехника, технологии беспроводной связи, большие данные и машинное обучение [8, 9]. При этом весьма непростой задачей для учителя является выбор инструментов, с помощью которых можно продемонстрировать возможности той или иной технологии [1, 2, 5, 3]. Для решения задач машинного обучения, таких, как задачи регрессии, классификации, кластеризации можно использовать возможности электронной таблицы.

Один из вариантов задачи анализа данных представлен в новой версии контрольно-измерительных материалов (задача №27) [4]. Задача 27 проверяет умение выполнять последовательность решения задач анализа данных: сбор первичных данных, очистку и оценку качества данных, выбор и построение модели, преобразование данных, интерпретацию результатов. Выпускникам предлагается решить задачу для двух наборов данных (файл А и файл Б), отличающихся объёмом: в файле А около 200 строк, в файле Б - около 20000 строк.

Для решения задачи можно использовать разные инструменты. В случае большого объёма данных (файл Б, 20000 строк) для решения задачи удобнее составить программу на языке программирования. Решение задачи в случае обработки данных меньшего объёма вполне можно выполнить с использованием электронной таблицы.

Задание №27 в рассматриваемой формулировке включено в содержание КИМ с 2025 года.

Методические материалы предназначены для подготовки к единому государственному экзамену (ЕГЭ) по информатике и включают разбор заданий разных типов и уровней сложности задания №27, а также подобраны задачи для самостоятельной работы, вызывающие наибольшую сложность у выпускников.

Задания взяты из открытого банка заданий ФИПИ.

1. **Спецификация задания**

Согласно спецификации контрольно-измерительных материалов для проведения в 2025 году основного государственного экзамена по информатике, задание №27 проверяет умение определять возможные результаты работы простейших алгоритмов управления исполнителями и вычислительных алгоритмов.

Код проверяемых элементов содержания (по кодификатору):

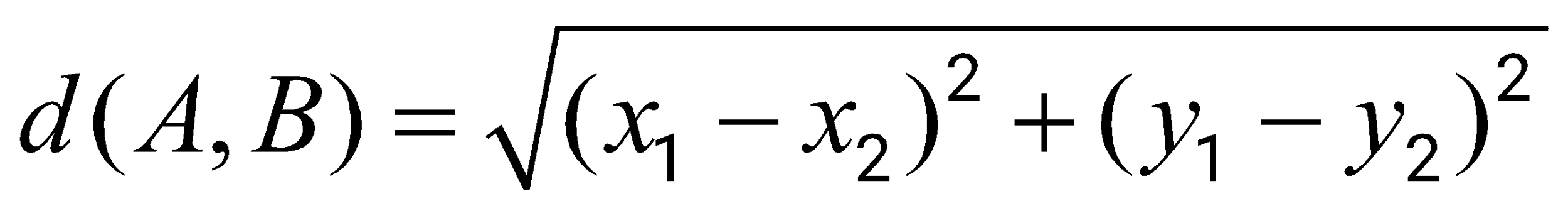
|  |  |  |  |
| --- | --- | --- | --- |
| **Код** | **Проверяемый элемент содержания** | **Уровень программы** | **Наличие данного элемента содержания в кодификаторе ОГЭ прошлых лет** |
| 4.1 | Анализ данных. Основные задачи анализа данных:  прогнозирование, классификация, кластеризация,  анализ отклонений. Последовательность решения  задач анализа данных: сбор первичных данных,  очистка и оценка качества данных, выбор и/или  построение модели, преобразование данных,  визуализация данных, интерпретация результатов.  Программные средства и интернет-сервисы для  обработки и представления данных. Большие дан-  ные. Машинное обучение | БУ, УУ | - |

Код проверяемых требований к уровню подготовки (по кодификатору):

|  |  |  |  |
| --- | --- | --- | --- |
| **Код проверяемого требования** | **Проверяемые требования к предметным результатам базового уровня освоения основной образовательной программы основного общего образования на основе ФГОС 2022 г.** | **Метапредметный результат** | **Обобщенные формулировки требований к предметным результатам из ФГОС 2012 г.** |
| 2.9 | Умение классифицировать основные задачи анализа  данных (прогнозирование, классификация, кластеризация, анализ отклонений); понимать последовательность решения задач анализа данных:  сбор первичных данных, очистка и оценка качества  данных, выбор и/или построение модели, пре-  образование данных, визуализация данных, интер-  претация результатов | МП 1.1 | Владение компьютерными средствами  представления и анализа данных (БУ);  наличие опыта использования компьютерных средств представления  и анализа данных (УУ) |

**Формулировка задания:**

*Фрагмент звездного неба спроецирован на плоскость с декартовой системой координат. Учёный решил провести кластеризацию полученных точек, являющихся изображениями звёзд, то есть разбить их множество на N непересекающихся непустых подмножеств (кластеров), таких, что точки каждого подмножества лежат внутри квадрата со стороной длиной Н, причём эти квадраты между собой не пересекаются. Стороны квадрата на обязательно параллельны координатным осям. Гарантируется, что такое разбиение существует и единственно для заданных размеров квадрата. некоторого множества звёзд по их расположению на карте звёздного неба. Кластер звёзд – это набор звёзд (точек) на графике, лежащий внутри прямоугольника высотой H и шириной W. Каждая звезда обязательно принадлежит только одному из кластеров.*

*Истинный центр кластера, или* ***центроид****, – это одна из звёзд на графике, сумма расстояний от которой до всех остальных звёзд кластера минимальна. Под расстоянием понимается расстояние Евклида между двумя точками A(x1, y1) и B(x2, y2) на плоскости, которое вычисляется по формуле:* *.*

***Входные данные***

*В файле A хранятся данные о звёздах двух кластеров, где H=3, W=3 для каждого кластера. В каждой строке записана информация о расположении на карте одной звезды: сначала координата x, затем координата y. Значения даны в условных единицах. Известно, что количество звёзд не превышает 1000.*

*В файле Б хранятся данные о звёздах трёх кластеров, где H=3, W=3 для каждого кластера. Известно, что количество звёзд не превышает 10 000. Структура хранения информации о звездах в файле Б аналогична файлу А.*

*Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров.*

***Выходные данные***

*В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 10 000, затем целую часть произведения Py × 10 000 для файла А, во второй строке – аналогичные данные для файла Б.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **№ зада-ния** | **Предметный результат обучения** | **Коды проверяемых элементов содержания по кодификатору** | **Коды требований к уровню подготовки по кодификатору** | **Уровень сложности задания** | **Максималь-ный балл за выполнение задания** | **Примерное время выполнения задания (мин.)** |
| 27 | Умение выполнять  последовательность  решения задач анализа  данных: сбор  первичных данных,  очистка и оценка  качества данных, выбор  и построение модели,  преобразование данных,  визуализация данных,  интерпретация  результатов | 4.1 | 2.2 | В | 2 | 40 |

1. **Результаты выполнения задания в 2024 году**

Задание включено в КИМ в 2025 году.

1. **Примеры заданий**

**4.1. Корреляционный анализ данных**

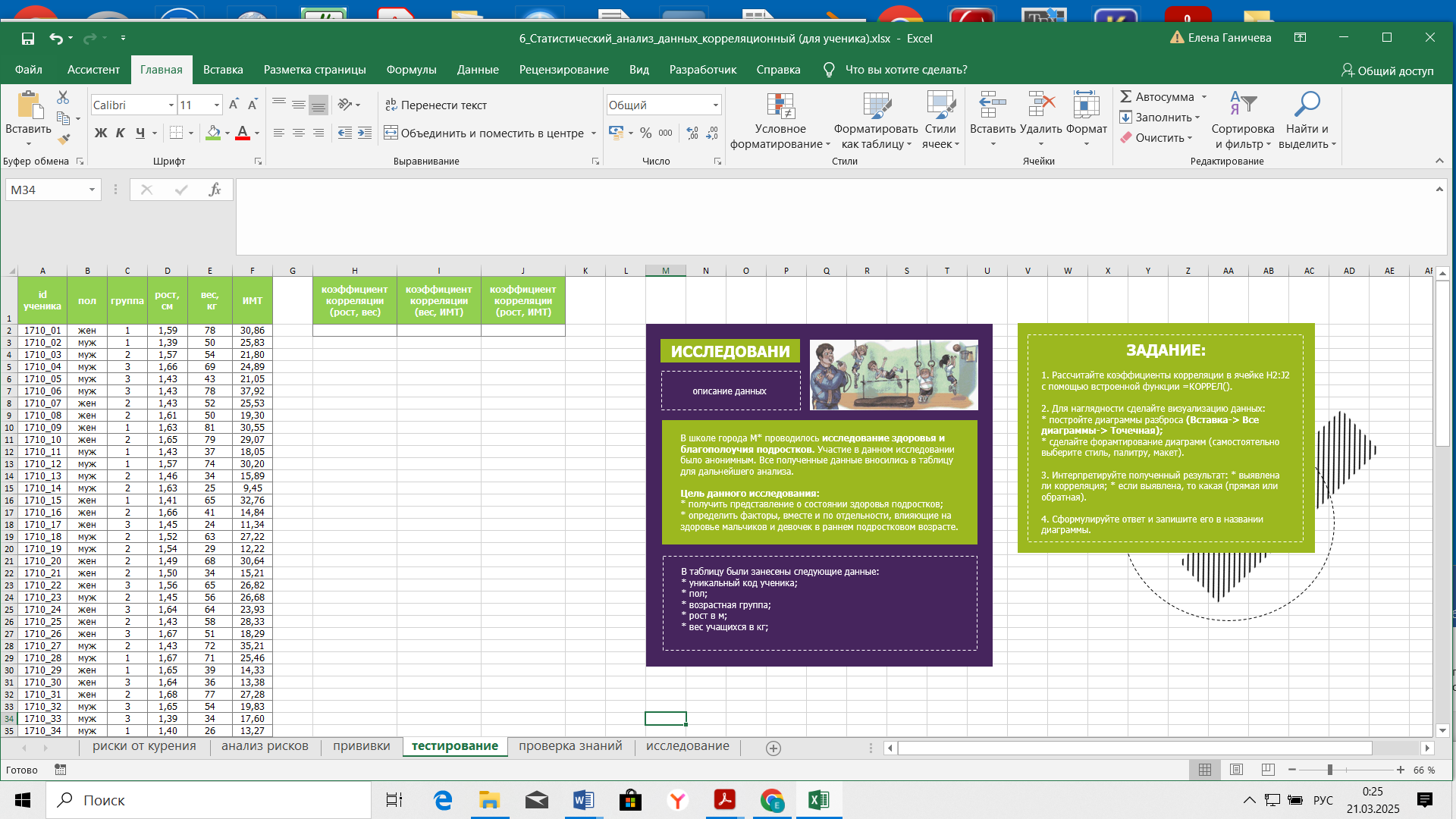
В научных и прикладных исследованиях часто требуется установить связь между результативными и факторными переменными (урожайностью какой-либо культуры и количеством осадков, ростом и весом человека в однородных группах по полу и возрасту, частотой пульса и температурой тела и т.д.). В этом помогает корреляционный анализ. Это метод обработки статистических данных, заключающийся в изучении коэффициентов корреляции между переменными. Для того, чтобы установить статистическую взаимосвязь между признаками, сравнивают коэффициенты корреляции между одной парой или множеством пар признаков.

**Задача 1.**

В школе города М\* проводилось исследование здоровья и благополоучия подростков. Участие в данном исследовании было анонимным. Все полученные данные вносились в таблицу (рисунок 1) для дальнейшего анализа.

Цель данного исследования:

* получить представление о состоянии здоровья подростков;
* определить факторы, вместе и по отдельности, влияющие на здоровье мальчиков и девочек в раннем подростковом возрасте.



*Рис. 1. Данные для задачи 1.*

**Задание:**

1. Рассчитайте коэффициенты корреляции в ячейке Н2:J2 с помощью встроенной функции =КОРРЕЛ().

2. Для наглядности сделайте визуализацию данных:

* постройте диаграммы разброса (Вставка → Все диаграммы → Точечная);
* сделайте форматирование диаграмм (самостоятельно выберите стиль, палитру, макет).

3. Интерпретируйте полученный результат: \* выявлена ли корреляция; \* если выявлена, то какая (прямая или обратная).

4. Сформулируйте ответ и запишите его в названии диаграммы.

**4.2. Линейный регрессионный анализ данных**

Корреляционный анализ позволяет выявить линейную связь между признаками и построить математическую модель линейной зависимости. В случае прямой или обратной корреляции точки на диаграмме образуют линию, похожую на прямую. Зная уравнение прямой и полученные параметры (коэффициент при **x** и свободный коэффициент), можно построить прямую, подставить интересующие нас **x** и **y** и сделать прогноз, предположить, что ждет нас в будущем.

**Цель регрессионного анализа** — предсказание ожидаемого среднего значения результирующей переменной с помощью уравнения регрессии. Например, анализ данных о продажах и закупках помогает выявить закономерности в покупках в определенные дни недели или время суток. Информация, полученная с помощью регрессионного анализа, помогает прогнозировать периоды времени, когда продукция компании будет пользоваться высоким спросом. Линейная регрессия важна и в спорте. Например, можно заинтересоваться

вопросом о том, зависит ли количество матчей, выигранных баскетбольной командой за сезон, от среднего количества очков за матч.

Уравнение линейной регрессии имеет вид, где: **y = ax + b, где** y — зависимая переменная;

x — независимая переменная; a и b — коэффициенты регрессии.

Коэффициент a характеризует наклон прямой, это — самый важный коэффициент. Коэффициент b определяет точку пересечения с осью OY (ордината).

Для вычисления коэффициентов регрессии в электронной таблице имеются встроенные функции **НАКЛОН** и **ОТРЕЗОК** :

Функция **НАКЛОН(известные\_значения\_y; исвестные\_значения\_x)** возвращает наклон линии линейной регрессии для точек данных в аргументах. Наклон определяется как частное от деления расстояния по вертикали на расстояние по горизонтали между двумя любыми точками прямой. Иными словами, наклон — это скорость изменения значений вдоль

прямой.

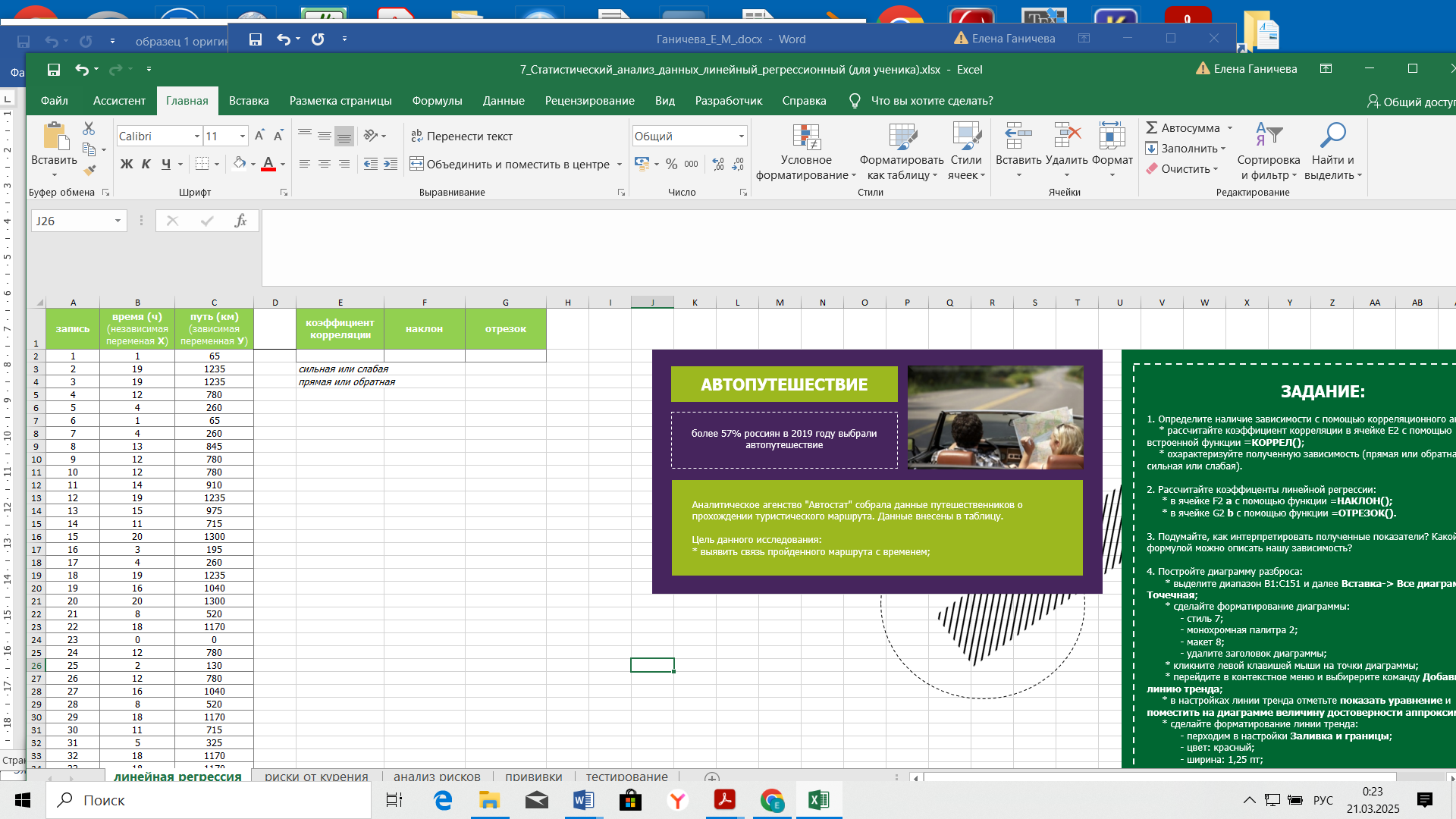
Функция **ОТРЕЗОК(известные\_значения\_y; известные\_значения\_x)** вычисляет точку пересечения линии с осью Y, используя значения аргументов. Точка пересечения находится на оптимальной линии регрессии, проведенной через точки, заданные аргументами. Функция используется для определения значения зависимой переменной при нулевом значении независимой переменной.

**Задача 2.**

Аналитическое агенство "Автостат" собрало данные путешественников о прохождении туристического маршрута. Данные внесены в таблицу (рисунок 2).

Цель данного исследования:

* выявить связь пройденного маршрута с временем;
* построить линейную математическую модель зависимости.



*Рис. 2. Данные для задачи 2.*

**Задание:**

1. Определите наличие зависимости с помощью корреляционного анализа:

* рассчитайте коэффициент корреляции в ячейке Е2 с помощью встроенной функции =КОРРЕЛ();
* охарактеризуйте полученную зависимость (прямая или обратная, сильная или слабая).

2. Рассчитайте коэффиценты линейной регрессии:

* в ячейке F2 а с помощью функции =НАКЛОН();
* в ячейке G2 b с помощью функции =ОТРЕЗОК().

3. Подумайте, как интерпретировать полученные показатели? Какой формулой можно описать нашу зависимость?

4. Постройте диаграмму разброса:

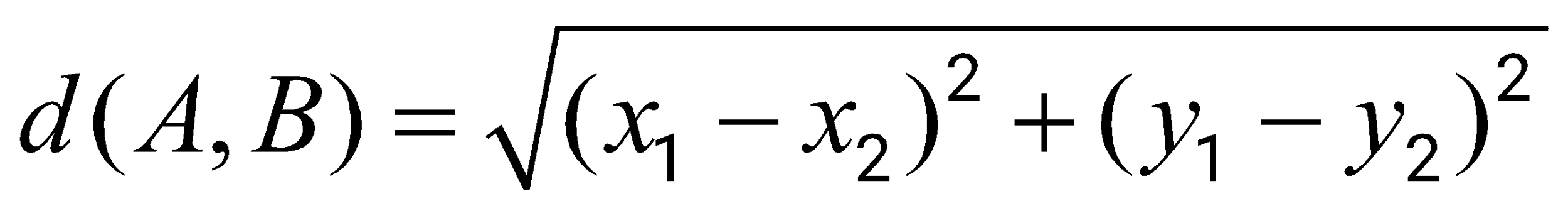
* постройте диаграммы разброса (Вставка → Все диаграммы → Точечная);
* сделайте форматирование диаграммы (самостоятельно выберите стиль, палитру, макет).
  + кликните левой клавишей мыши на точки диаграммы;
  + перейдите в контекстное меню и выберите команду Добавить линию тренда;
  + в настройках линии тренда отметьте показать уравнение и поместить на диаграмме величину достоверности аппроксимации.

**4.3. Задача кластеризации, решаемая в табличном процессоре**

В задачах данного типа необходимо определить количество кластеров, найти центроид в каждом кластере и дать ответ на поставленный в задаче вопрос, выполнив действия с координатами кластеров.

Само задание формулируется следующим образом:

*Фрагмент звездного неба спроецирован на плоскость с декартовой системой координат. Учёный решил провести кластеризацию полученных точек, являющихся изображениями звёзд, то есть разбить их множество на N непересекающихся непустых подмножеств (кластеров), таких, что точки каждого подмножества лежат внутри квадрата со стороной длиной Н, причём эти квадраты между собой не пересекаются. Стороны квадрата на обязательно параллельны координатным осям. Гарантируется, что такое разбиение существует и единственно для заданных размеров квадрата. некоторого множества звёзд по их расположению на карте звёздного неба. Кластер звёзд – это набор звёзд (точек) на графике, лежащий внутри прямоугольника высотой H и шириной W. Каждая звезда обязательно принадлежит только одному из кластеров.*

*Истинный центр кластера, или* ***центроид****, – это одна из звёзд на графике, сумма расстояний от которой до всех остальных звёзд кластера минимальна. Под расстоянием понимается расстояние Евклида между двумя точками A(x1, y1) и B(x2, y2) на плоскости, которое вычисляется по формуле:* *.*

***Входные данные***

*В файле A хранятся данные о звёздах двух кластеров, где H=3, W=3 для каждого кластера. В каждой строке записана информация о расположении на карте одной звезды: сначала координата x, затем координата y. Значения даны в условных единицах. Известно, что количество звёзд не превышает 1000.*

*В файле Б хранятся данные о звёздах трёх кластеров, где H=3, W=3 для каждого кластера. Известно, что количество звёзд не превышает 10 000. Структура хранения информации о звездах в файле Б аналогична файлу А.*

*Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров.*

***Выходные данные***

*В ответе запишите четыре числа: в первой строке сначала целую часть произведения Px × 10 000, затем целую часть произведения Py × 10 000 для файла А, во второй строке – аналогичные данные для файла Б.*

Задача может быть решена на языке программирования как в случае А, так и в случае Б. Однако для многих обучающихся сложно понять суть задачи, поэтому полезно рассмотреть способ решения задачи с использованием данных из файла А. в электронной таблице.

**Решение.**

**Идея решения:**

1. Выполним визуализацию данных, построив точечную диаграмму.
2. По внешнему виду диаграммы определим условия разделения кластеров.
3. Выполним сортировку данных согласно условию разделения так, чтобы в верхней части диапазона находились строки с данными, для которых выполнено условие разделения.
4. Для каждого из кластеров создадим новый лист электронной таблицы, перенесем на него строки с координатами точек, относящихся к этому кластеру и подготовим таблицы для определения центроида.
5. Определим координаты центроидов всех кластеров.
6. Выполним действия с координатами.

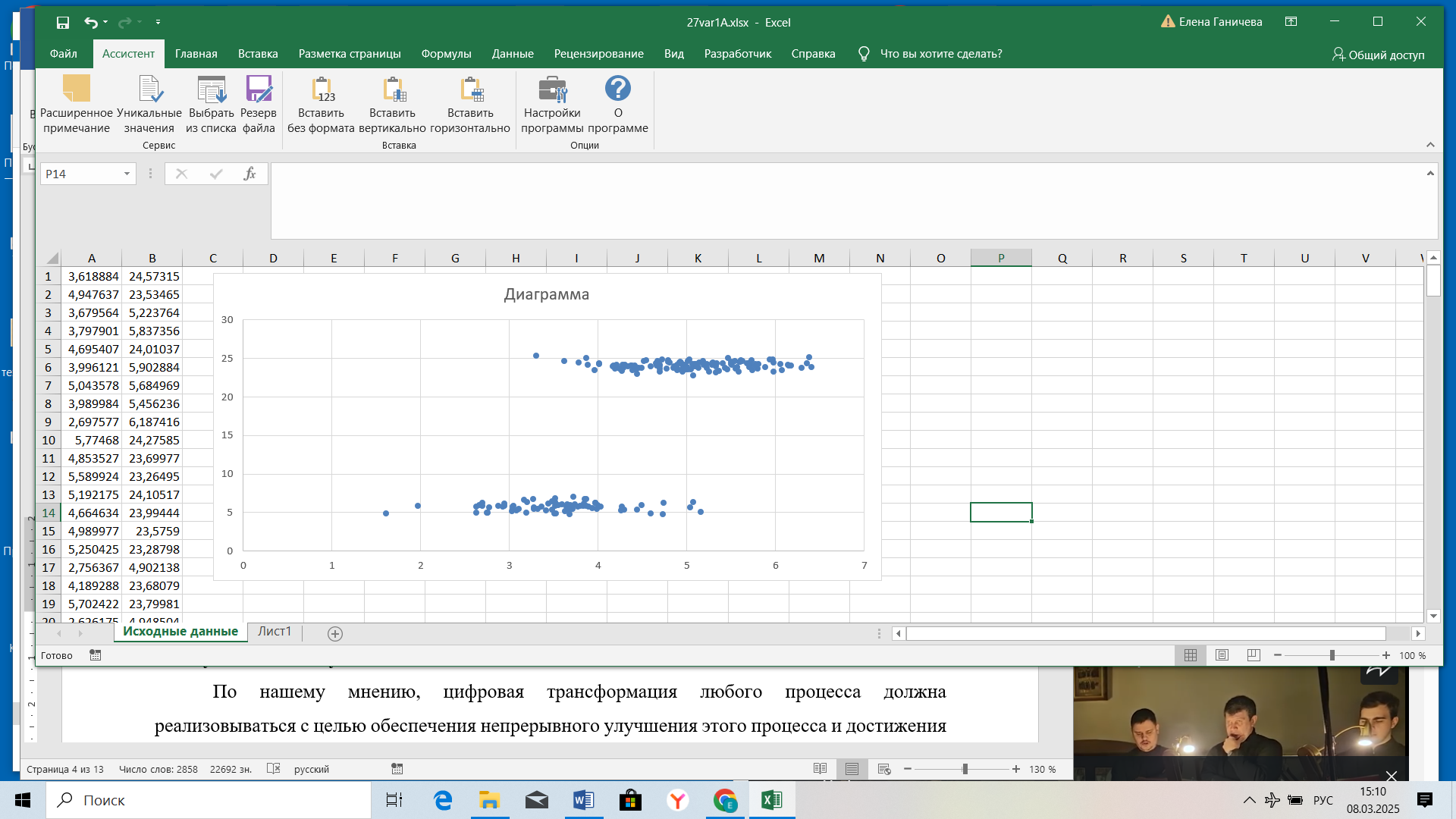
**Описание алгоритма решения задачи:**

Исходные данные о координатах точек (звёзд) представлены в файле электронной таблицы в виде двух столбцов данных. Для решения задачи надо определить количество кластеров. Это удобно сделать, построив точечную диаграмму.

**Способ для электронной таблицы Excel:** выделить данные (диапазон A1: B201) – Вставка – Диаграмма – Точечная.

**Способ для электронной таблицы Libre Calc:** выделить данные (диапазон A1: B201) – Вставка – Диаграмма – Х-У разброс.

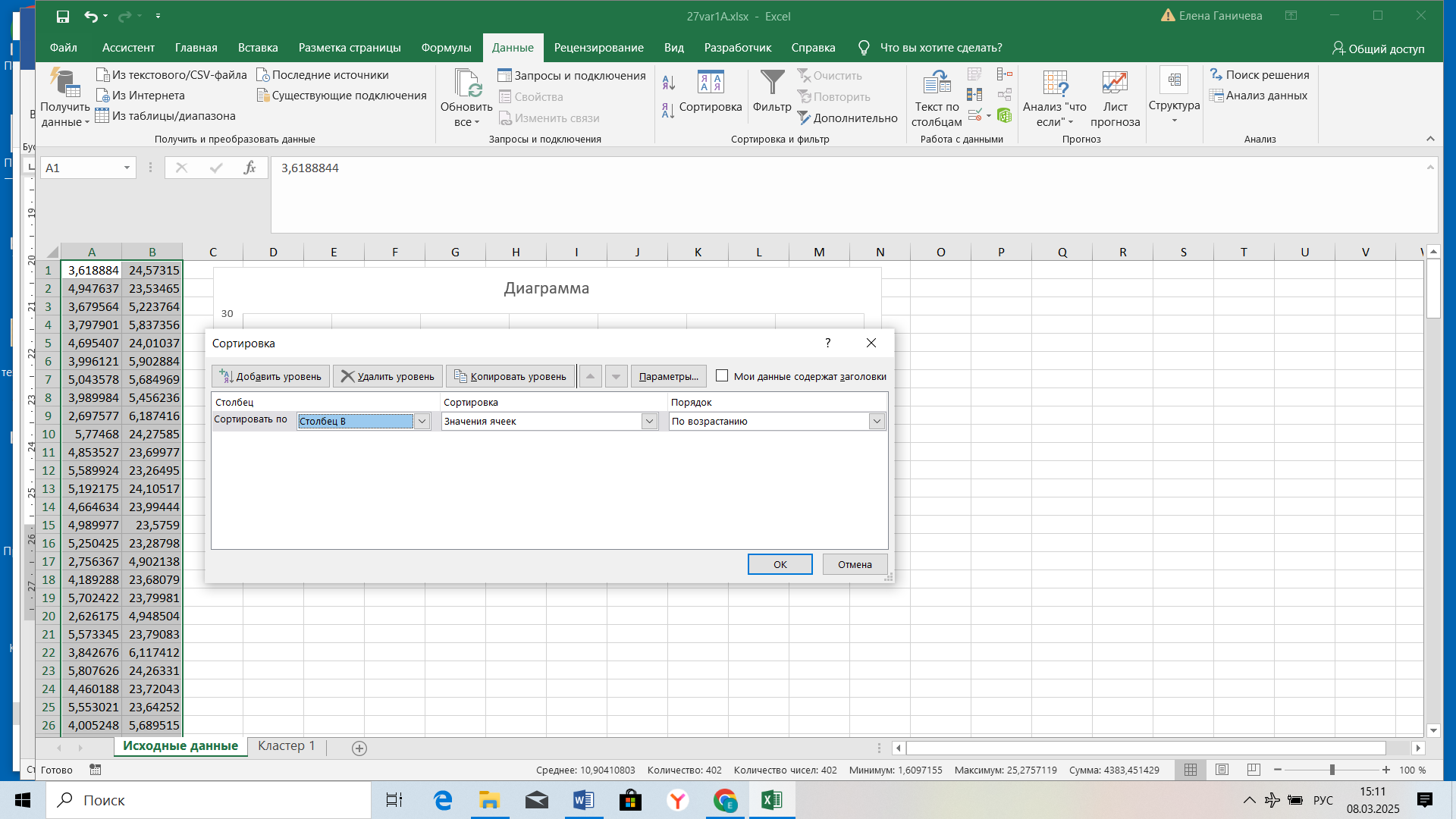
В результате получим диаграмму, представленную на рисунке 3:



*Рис. 3. Диаграмма точек кластеров*

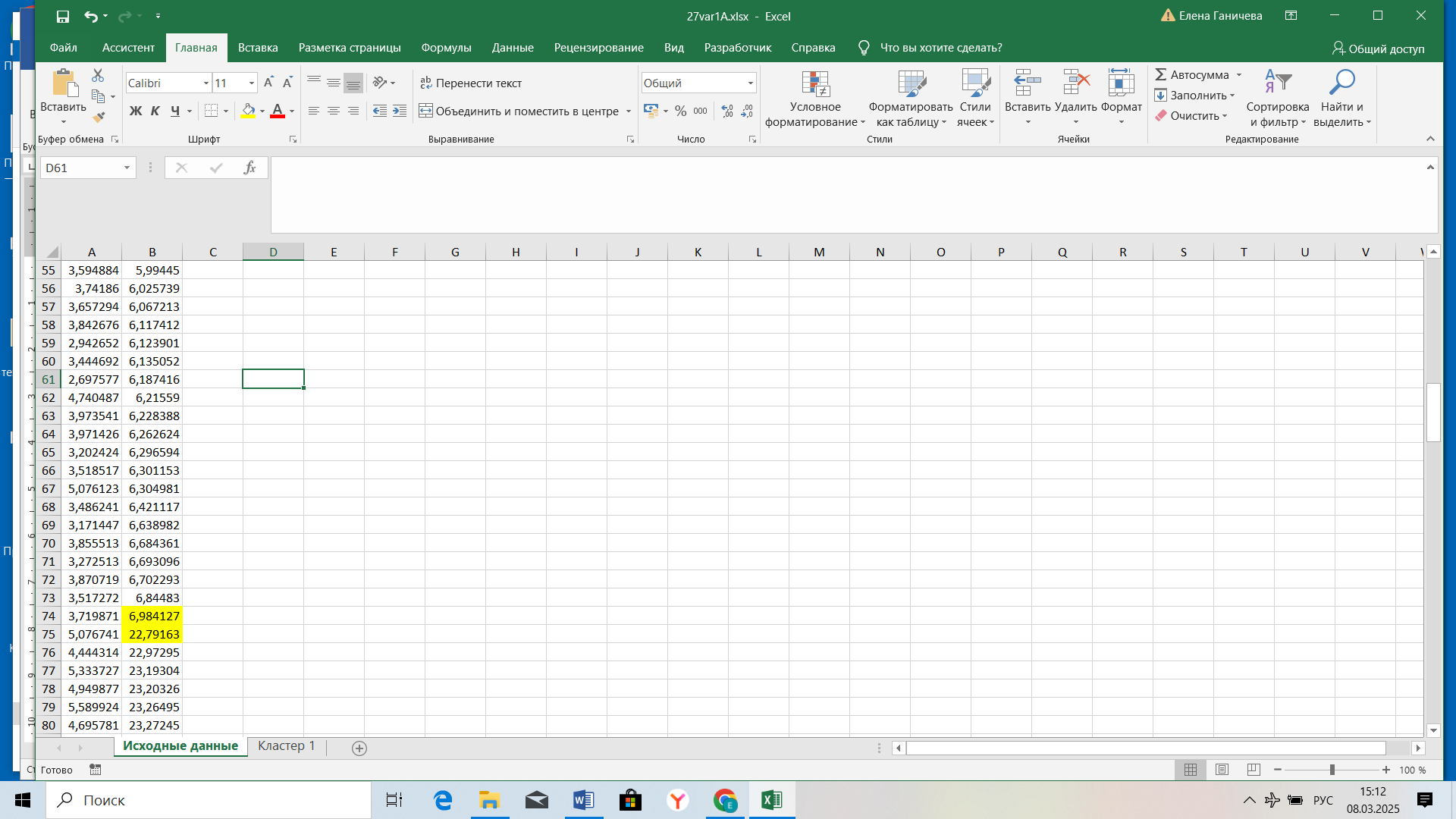
На рисунке явно виды два кластера точек. Надо определить, как задать границу разделения. В нашем примере такой границей может стать прямая линия, заданная, например, уравнением у=10. Назовем кластер, расположенный в нижней части диаграммы, первым, а кластер, расположенный выше, вторым.

Для того, чтобы найти координаты центроида, необходимо разделить массив значений координат точек на два множества: точки верхнего и точки нижнего кластеров. Это можно сделать, выполнив сортировку данных по второму столбцу (координата У) по возрастанию (рисунок 4).



*Рис. 4. Окно параметров сортировки*

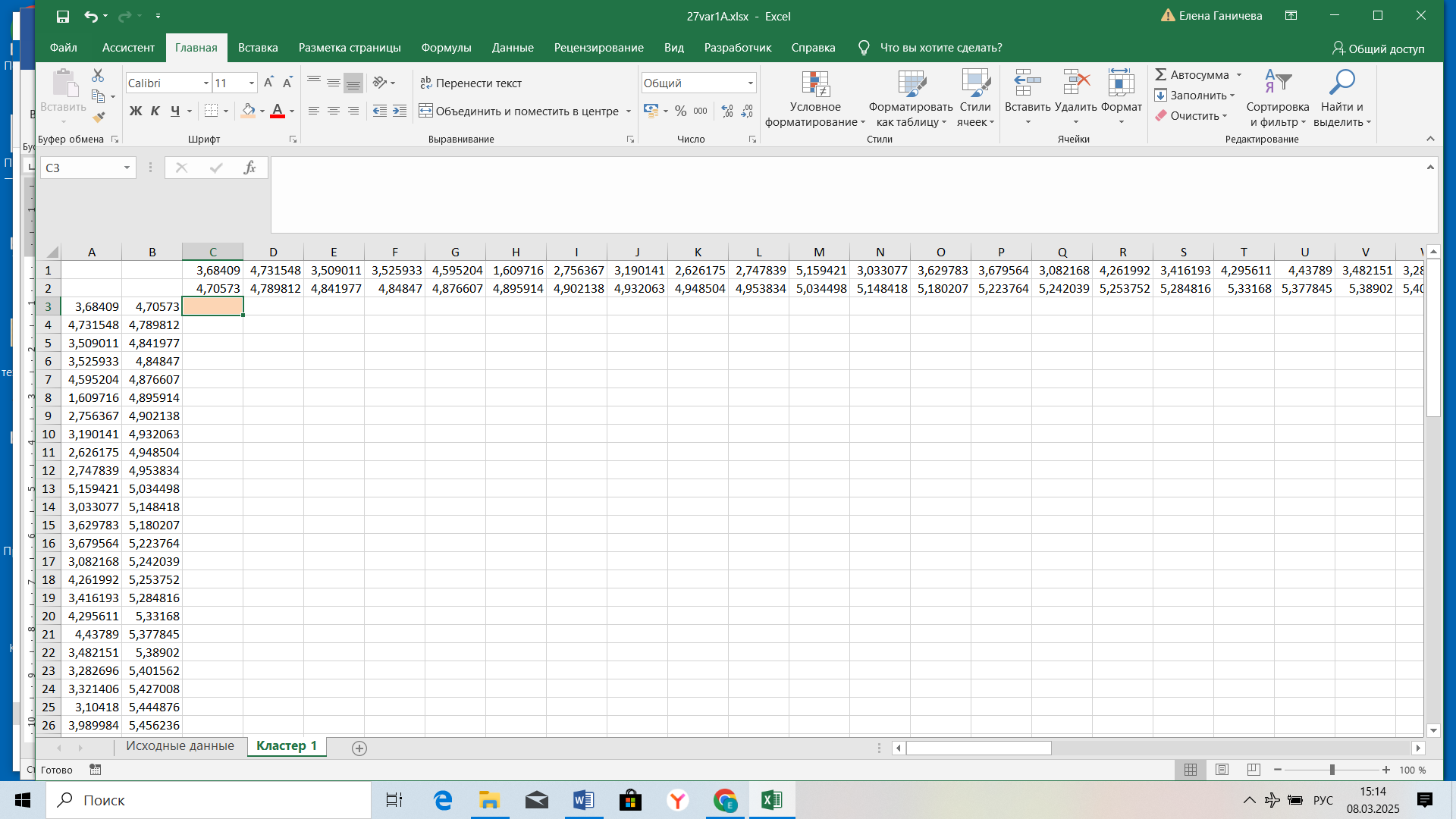
После сортировки точки, координаты у которых имеют значения, меньшие 10, будут расположены в строках с 1 по 74. Заметим, что в 75-й строке координата у будет больше 10 (рисунок 5):



*Рис. 5. Фрагмент результата сортировки*

Для вычисления расстояний между точками первого кластера добавим новый лист, переименуем его в Кластер 1 и скопируем данные строк с 1 по 74 на лист Кластер 1. Вставим значения в ячейки диапазона А3:В76.

Этот же массив данных вставим, начиная с ячейки С1, с транспонированием. (Транспонирование – это изменение расположения элементов: столбец станет строкой) (рисунок 6).

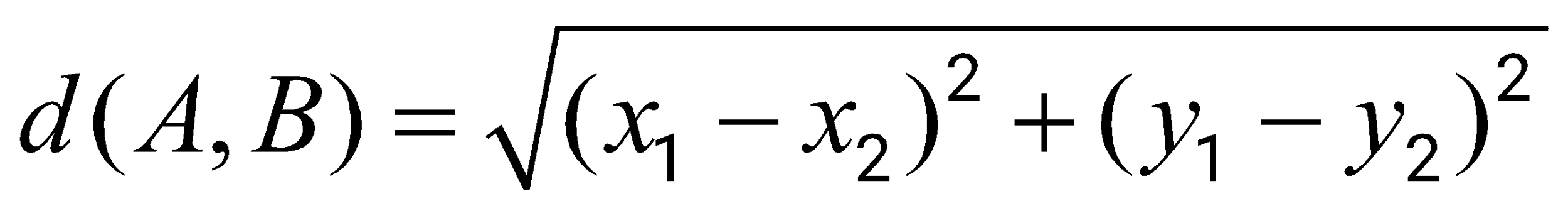


*Рис. 6. Результат вставки данных на лист Кластер 1*

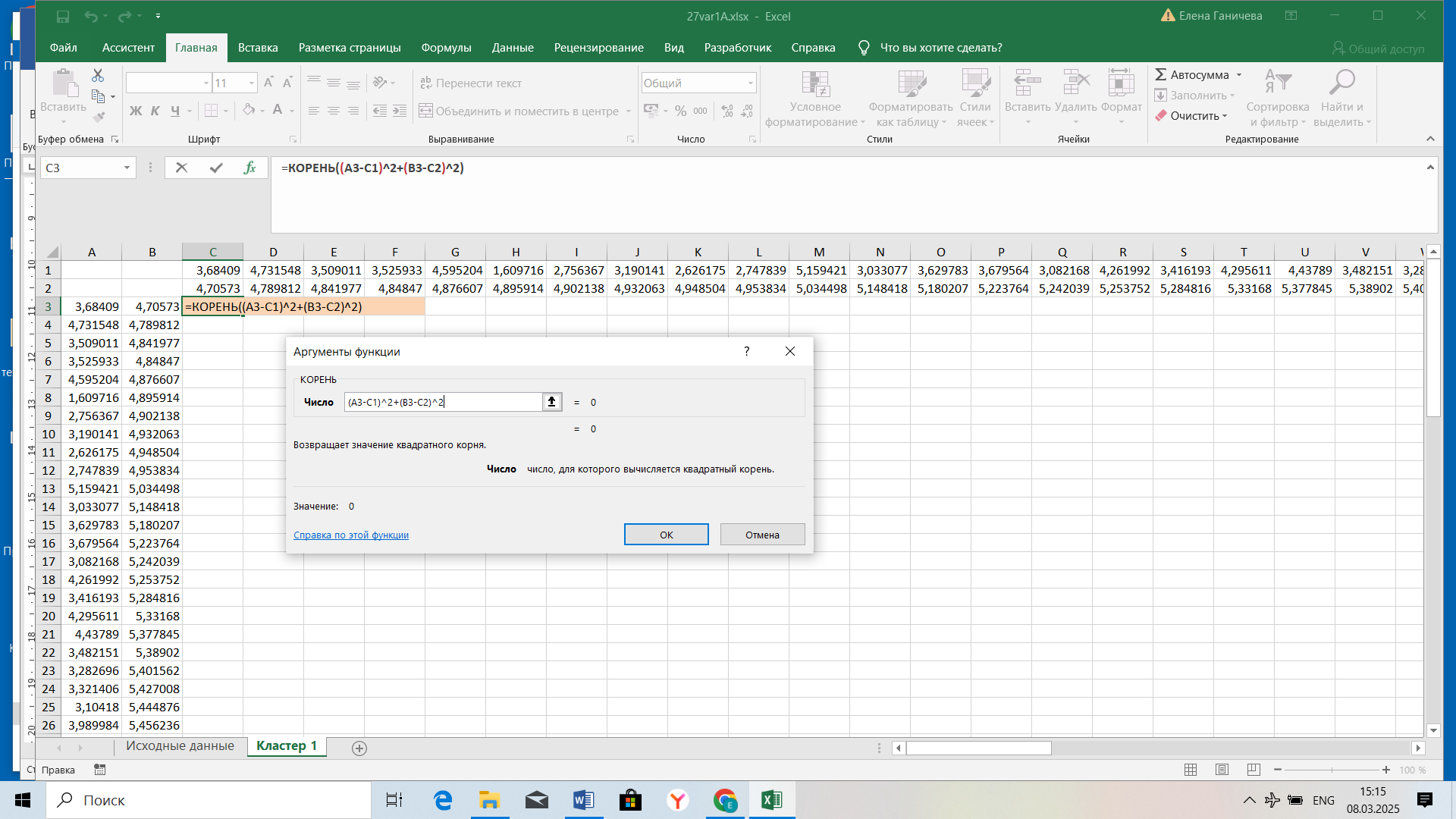
Как выполнить вставку с транспонированием элементов столбца? **Способ для электронной таблицы Excel:** выделить данные – копировать – установить курсор в ячейку С1 – Вставка – с транспонированием.

**Для электронной таблицы Libre Calc:** выделить данные – копировать – установить курсор в ячейку С1 – Правка – Вставить как – транспонировать. При пояснении этих действий можно привести аналогию, например, с таблицей умножения или таблицей квадратов чисел.

В ячейку С3 запишем формулу для вычисления расстояния между точками:

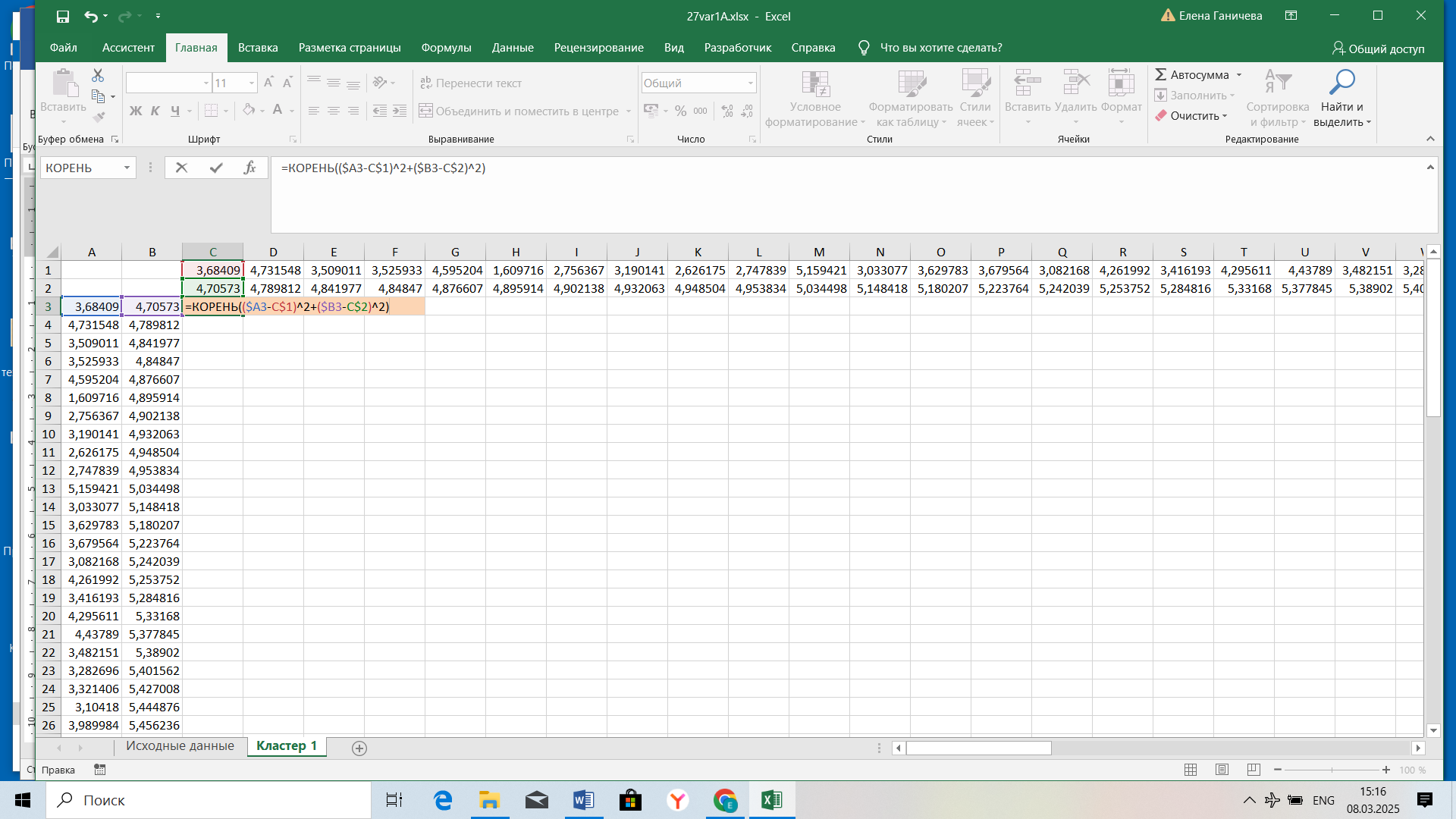


При этом координаты точки (х1, у1) находятся в ячейках А3 и В3, координаты точки (х2, у2) находятся в ячейках С1 и С2 (рисунок 7).



*Рис. 7. Окно аргументов функции*

Для того, чтобы формулу можно было скопировать на весь диапазон, зададим абсолютную адресацию. Необходимо зафиксировать имена столбцов А и В и номера строк 1 и 2. Формула примет вид, как на рисунке 8:

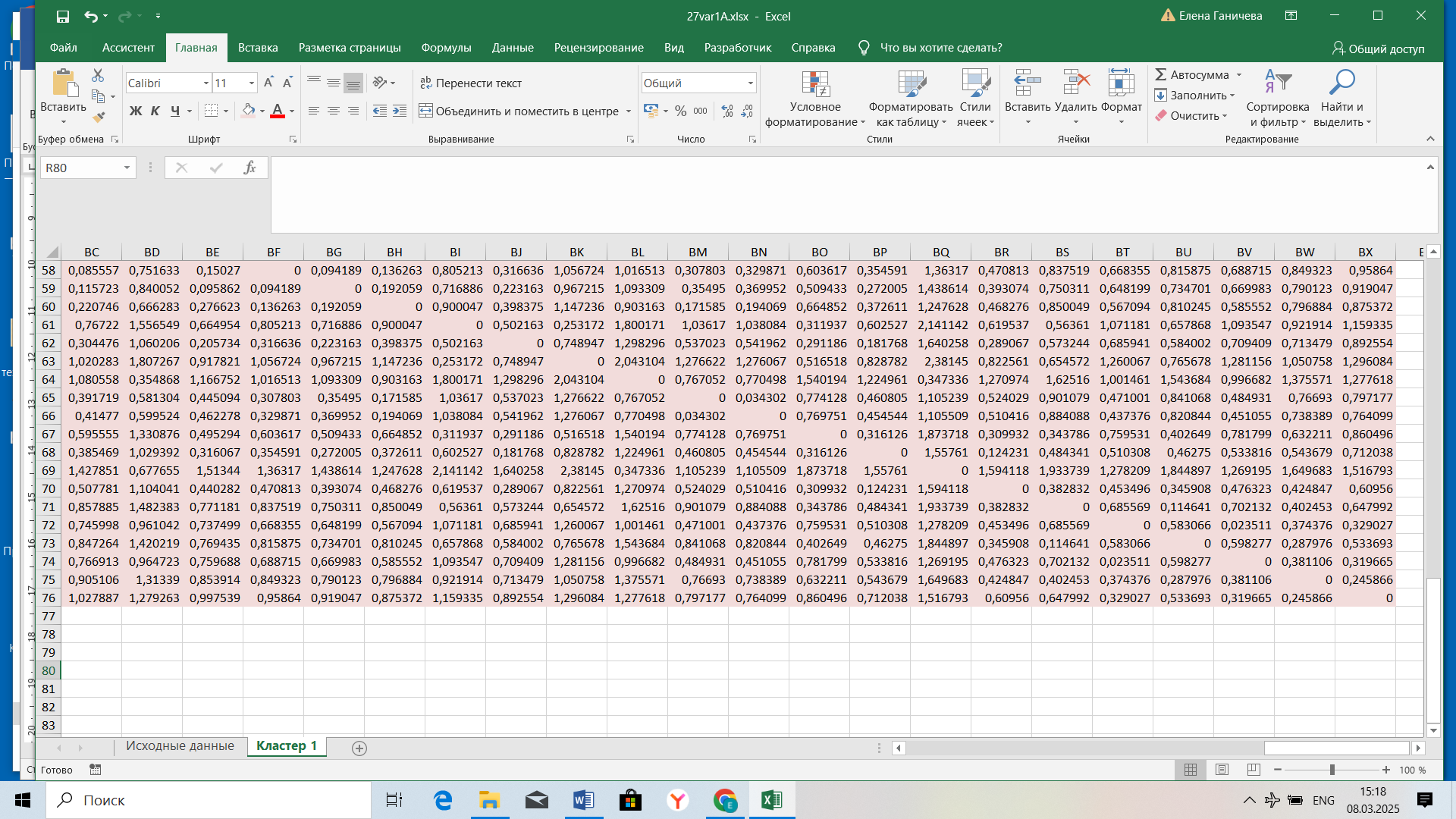


*Рис. 8. Формула для вычисления расстояния*

Значения в ячейке С3 будет равно нулю, а в остальных ячейках первой строки получим расстояния от первой точки до всех 73-х точек кластера. Во второй строке – расстояния от второй до всех остальных точек кластера и т.д.

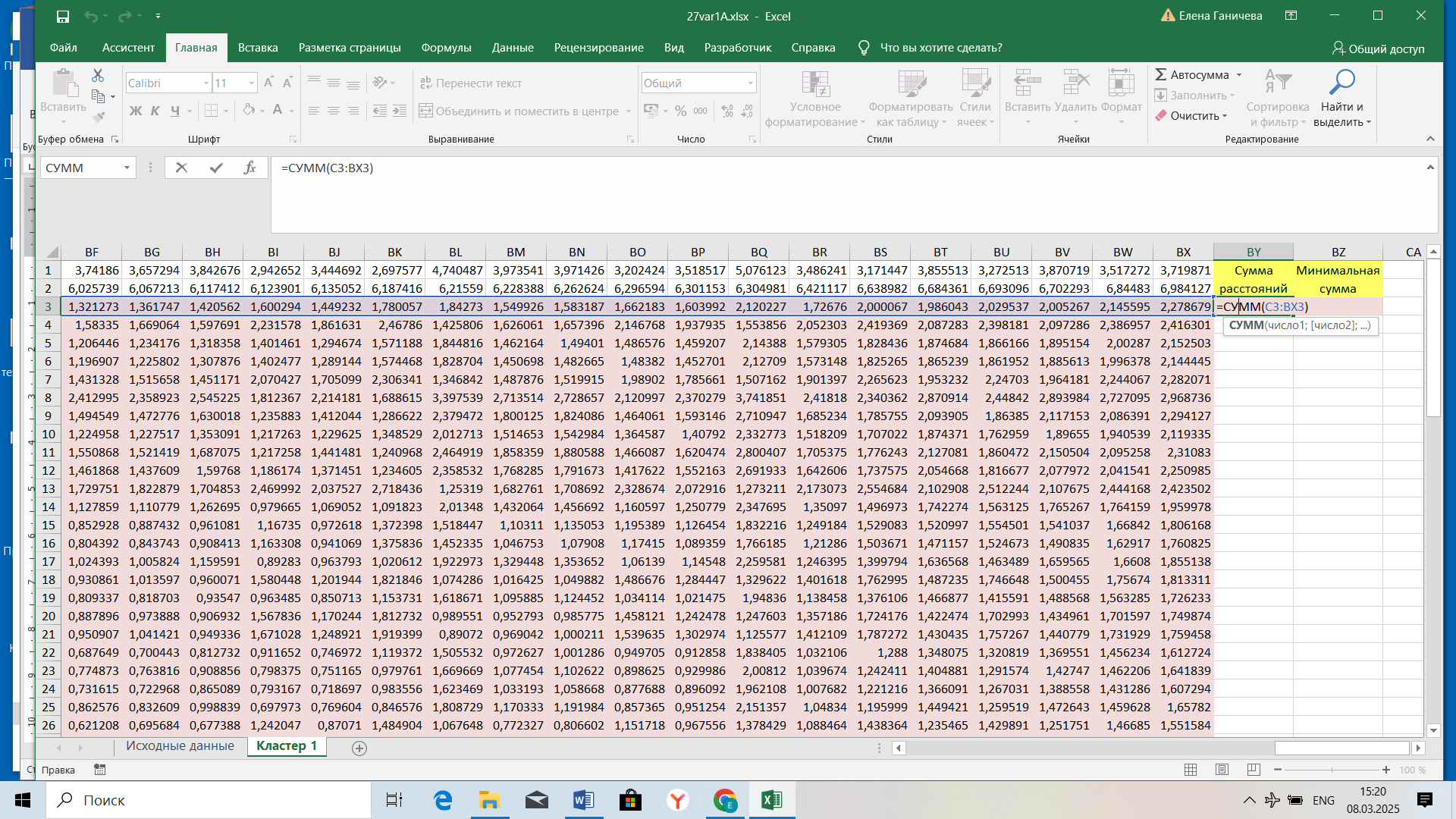
Удобно скопировать формулу из ячейки С3 во все ячейки столбца С3:С76, а затем продолжить копирование всего столбца по горизонтали.

При этом нули на диагонали помогут ориентироваться при определении диапазона копирования. Нуль должен оказаться в правом нижнем углу диапазона (рисунок 9).



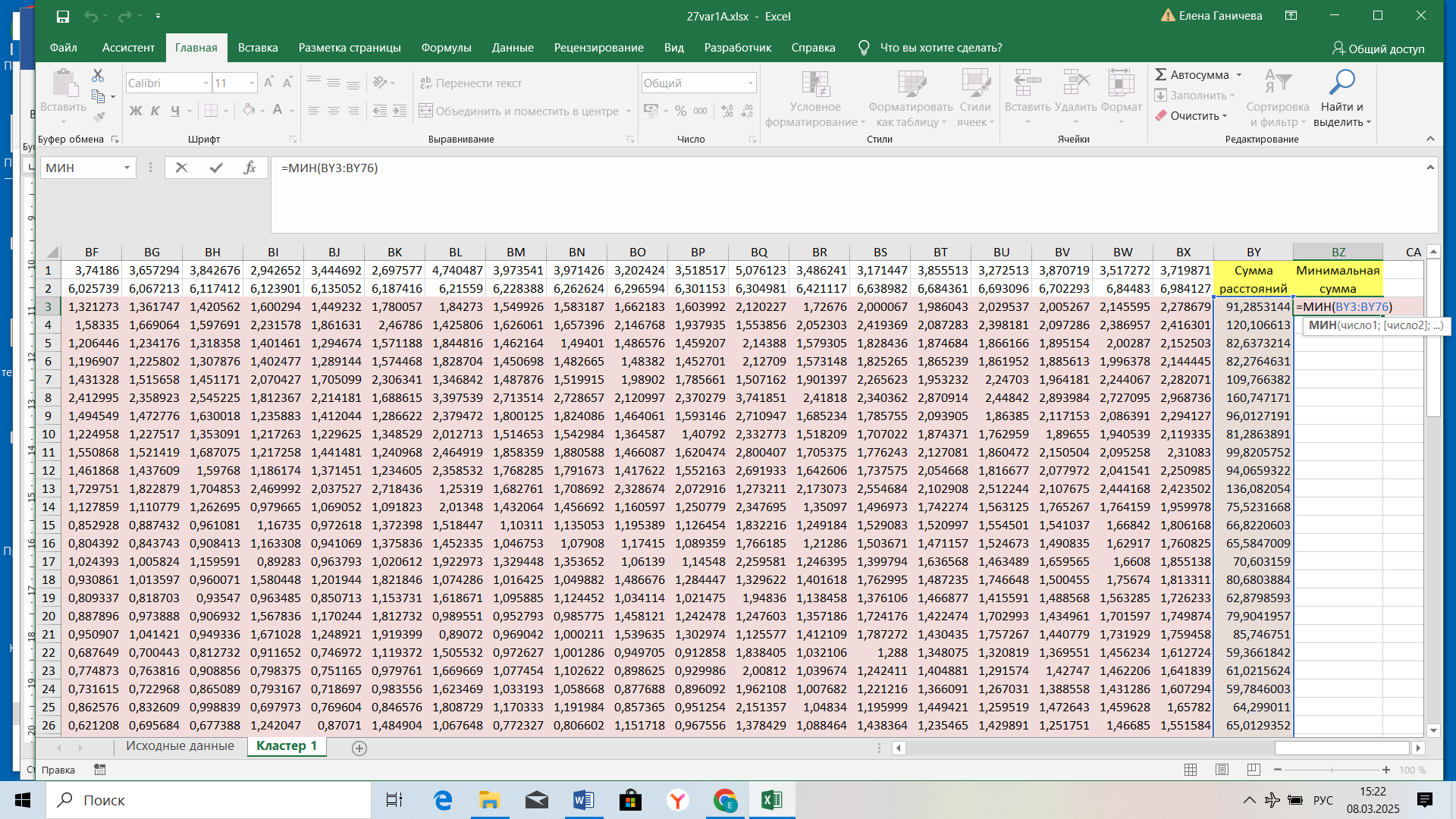
*Рис. 9. Результат копирования формулы для вычисления расстояния*

Следующим шагом станет вычисление суммы расстояний по строке (рисунок 10):



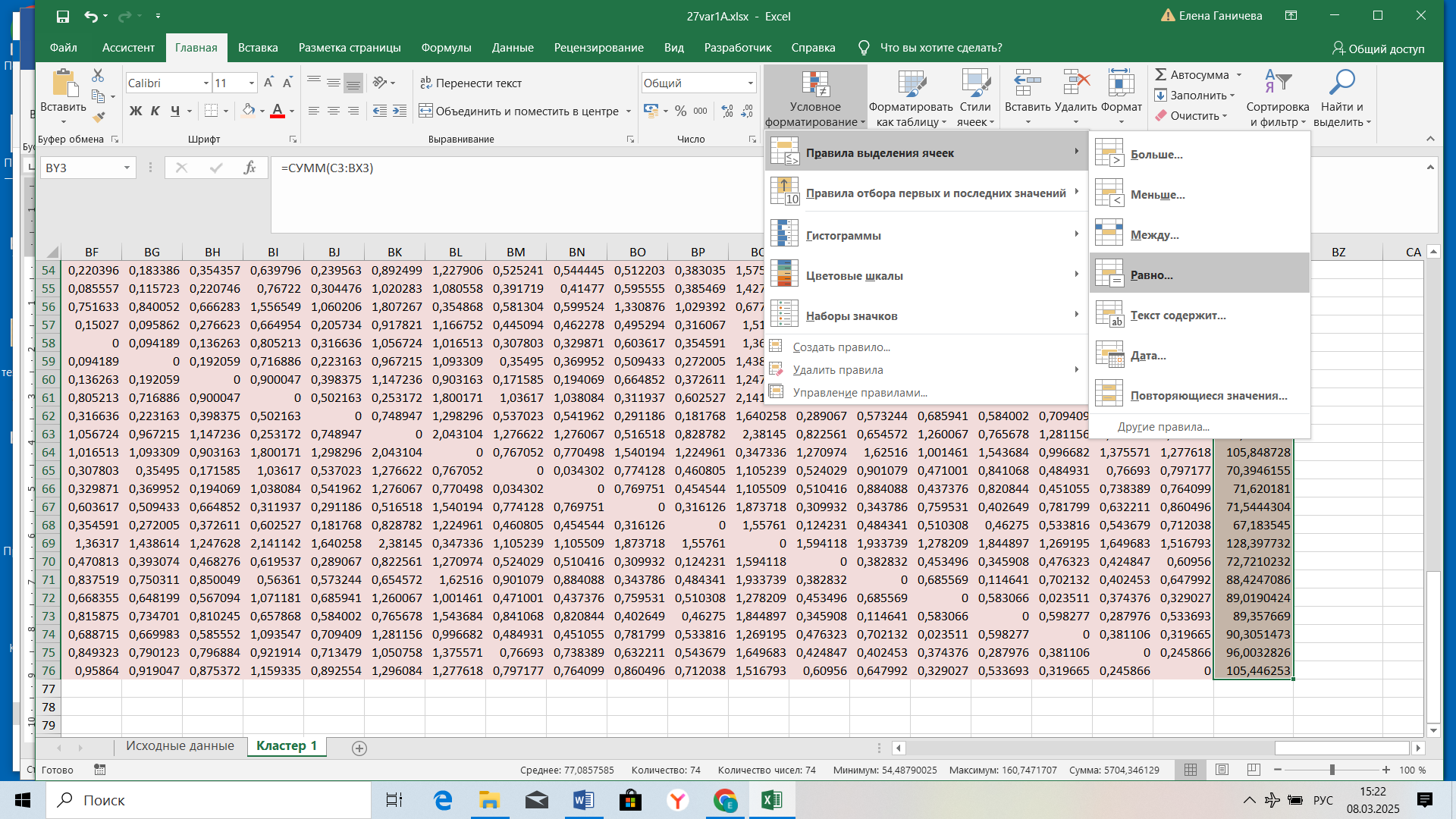
*Рис. 10. Формула для вычисления суммы расстояний*

Далее находим минимальную сумму расстояний от одной точки до всех остальных точек кластера (рисунок 11):



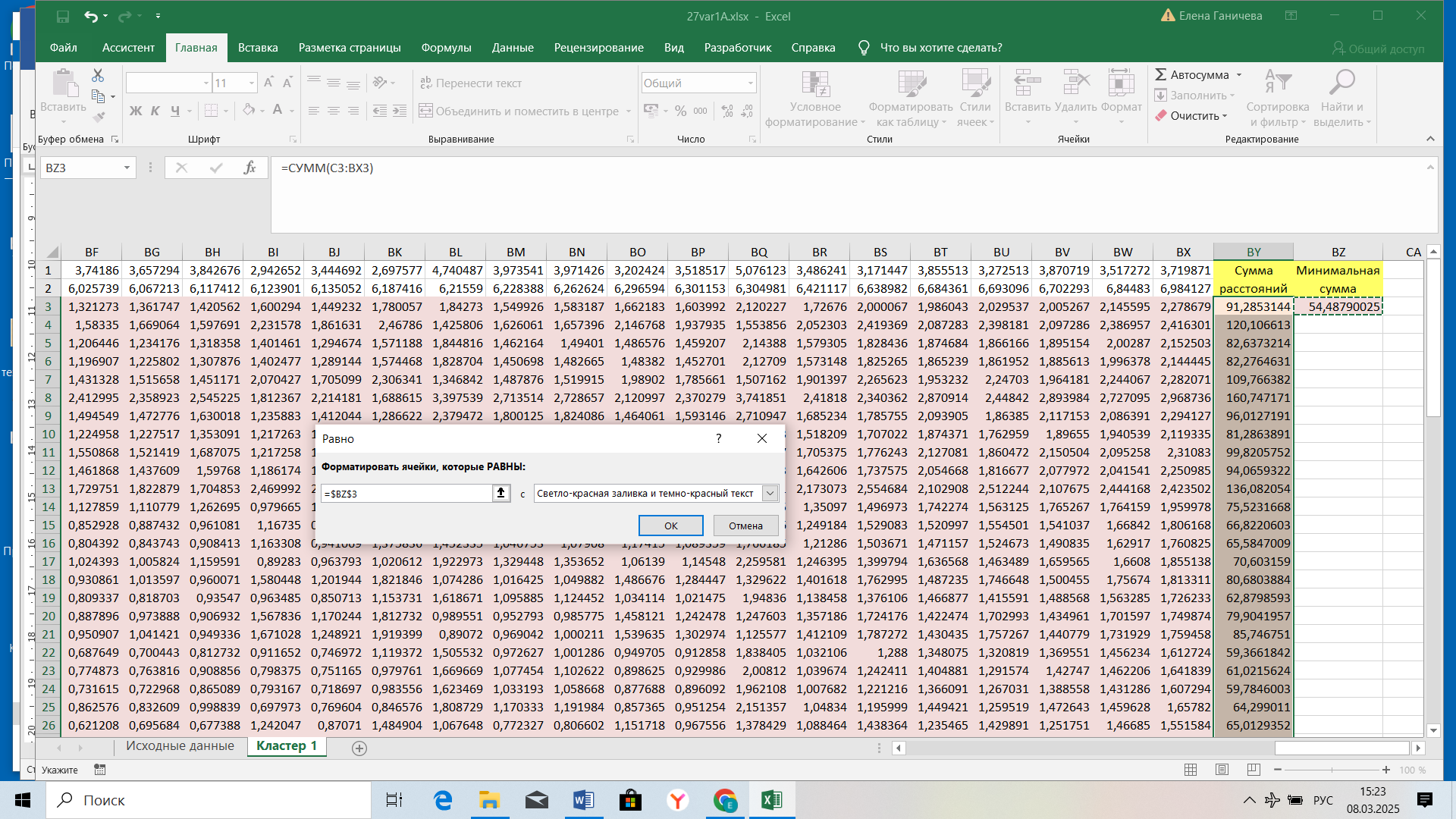
*Рис. 11. Формула для определения минимальной суммы расстояний*

Поиск нужного значения в столбце BY можно выполнить разными способами: визуально, применить средства поиска или условное форматирование (рисунок 12):



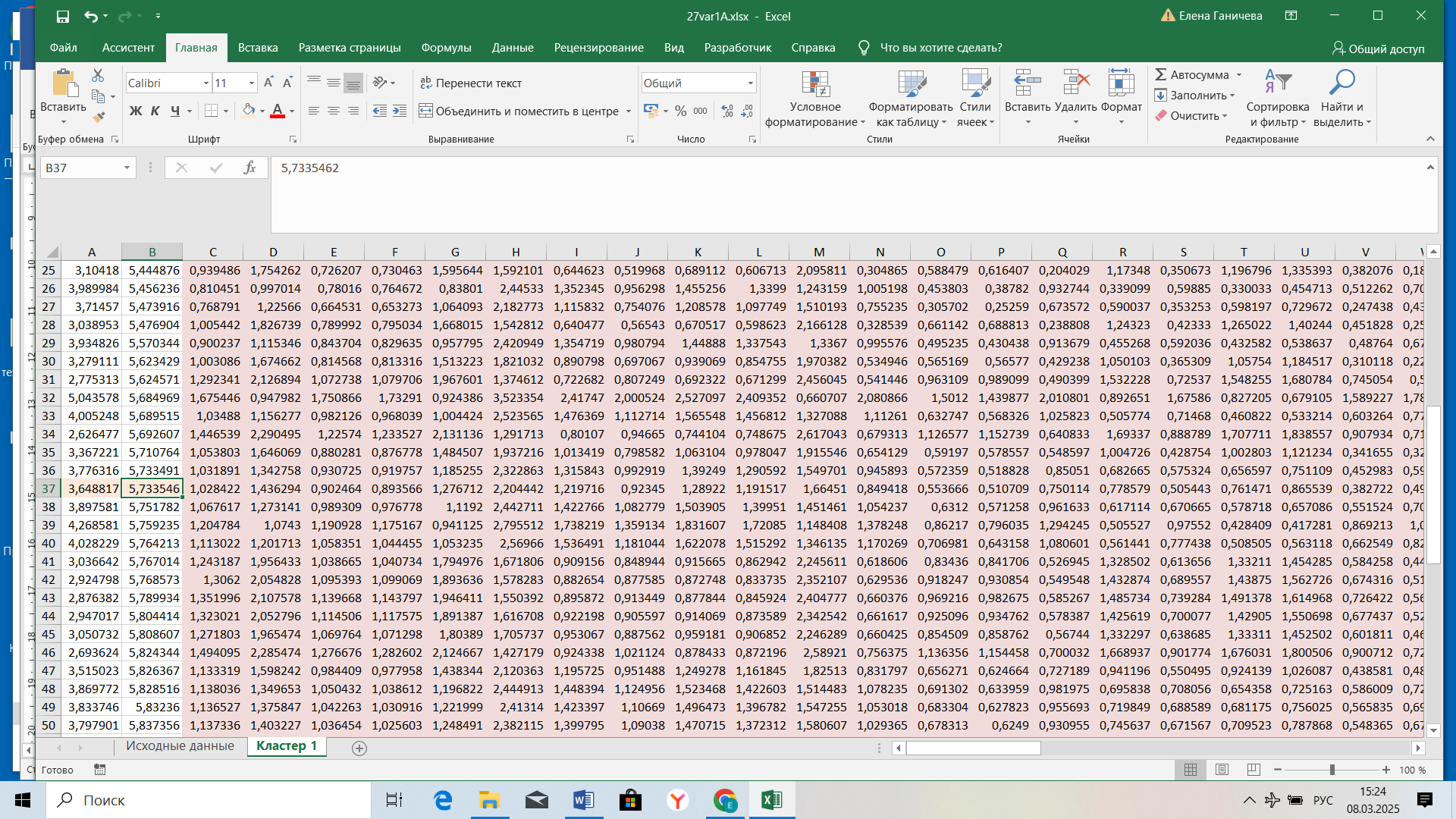
*Рис. 12. Панель «Условное форматирование»*

Правило выделения ячеек задаем равным полученному минимальному значению суммы расстояний (рисунок 13):



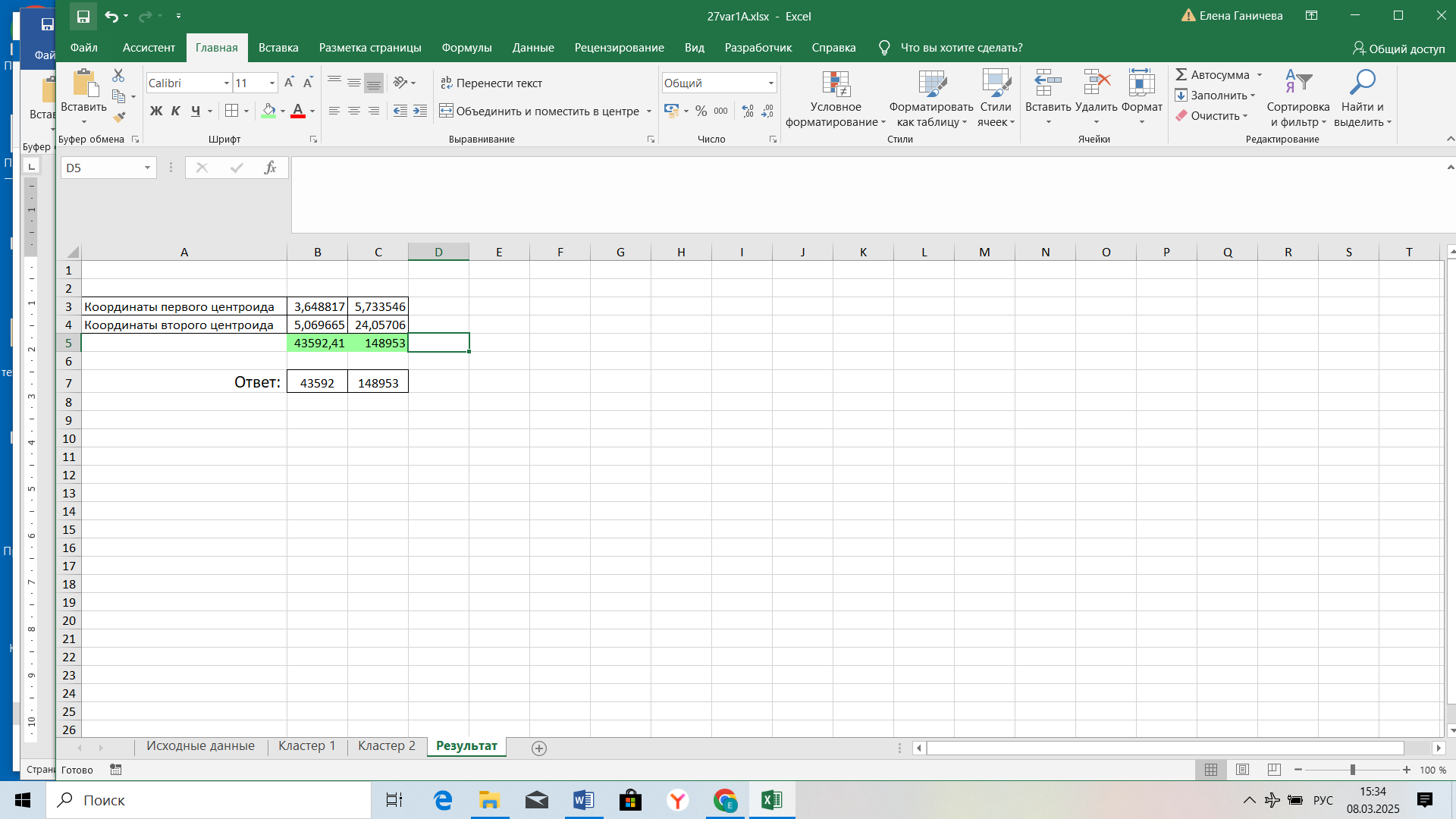
*Рис. 13. Окно аргументов условного форматирования*

После нахождения нужного значения в столбце BY по строке определим координаты центроида. В нашем случае координаты центроида находятся в 37-й строке (рисунок 14).



*Рис. 14. Координаты центроида первого кластера*

Добавим новый лист Результат и вставим найденные значения координат. Координаты второго центроида найдем по такому же алгоритму и вставим найденные значения на лист Результат. Вычислим Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров, умножим полученные значения на 10000 и запишем ответ (рисунок 15):



*Рис. 15. Результат решения задачи*

Электронную таблицу можно применить и для визуализации данных, представленных в файле Б, что является необходимым для определения количества кластеров и условий принадлежности точек к конкретному кластеру.

1. **Разбор ключевых задач**

**Задача 1.** Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба. Кластер звёзд – это набор звёзд (точек) на графике, лежащий внутри прямоугольника высотой H и шириной W. Каждая звезда обязательно принадлежит только одному из кластеров.

Истинный центр кластера, или **центроид**, – это одна из звёзд на графике, сумма расстояний от которой до всех остальных звёзд кластера минимальна. Под расстоянием понимается расстояние Евклида между двумя точками A(x1, y1) и B(x2, y2) на плоскости, которое вычисляется по формуле: .

**Входные данные**

В файле A хранятся данные о звёздах двух кластеров, где H=3, W=3 для каждого кластера. В каждой строке записана информация о расположении на карте одной звезды: сначала координата x, затем координата y. Значения даны в условных единицах. Известно, что количество звёзд не превышает 1000.

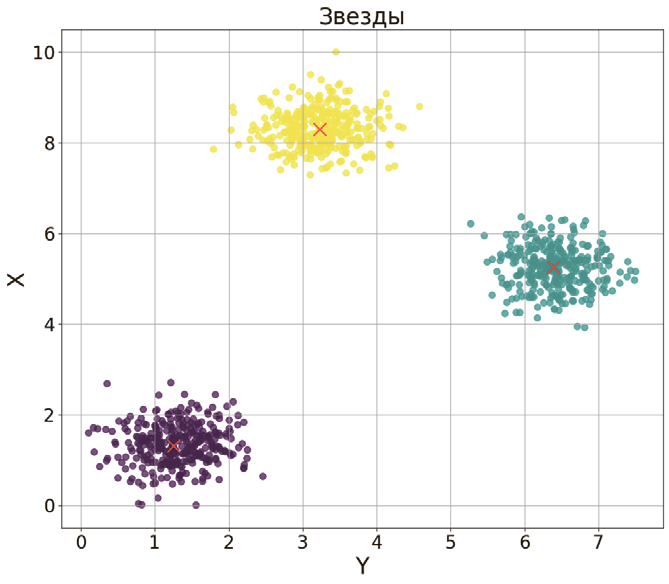
В файле Б хранятся данные о звёздах трёх кластеров, где H=3, W=3 для каждого кластера. Известно, что количество звёзд не превышает 10 000. Структура хранения информации о звездах в файле Б аналогична файлу А.

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: Px – среднее арифметическое абсцисс центров кластеров, и Py – среднее арифметическое ординат центров кластеров.

**Выходные данные**

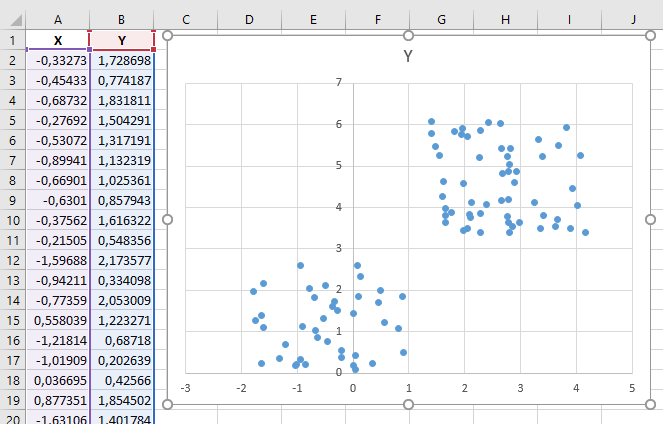
В ответе запишите четыре числа: в первой строке сначала целую часть абсолютного значения произведения Px × 10 000, затем целую часть абсолютного значения произведения Py × 10 000 для файла А, во второй строке – аналогичные данные для файла Б.

Возможные данные одного из файлов иллюстрированы графиком.



**Решение:**

1. Для каждой задачи (А и Б) даны два файла: один текстовый, второй – с теми же данными в виде электронной таблицы. Используя электронную таблицу, мы можем сразу определить границы кластеров (визуально).
2. Откроем электронную таблицу для задачи А и построим точечную диаграмму.



Видно, что здесь можно выделить два кластера, которые разделяются прямой *x* = 1. Все точки слева от этой прямой нужно отнести к кластеру 0 (в программе нам удобнее сделать нумерацию кластеров с нуля), а все точки с *x* > 1 – к кластеру 1. Напишем функцию, которая будет определяет номер кластера по координатам:

**def** findClusterNo( x, y ):

**return** 0 **if** x < 1 **else** 1

1. Обозначим через K количество кластеров (K = 2 для задачи А и K = 3 для задачи Б).

K = 2

1. Данные о кластерах будет хранить в списке:

clusters = [ [], [] ]

# или clusters = [ [] **for** i **in** range(K) ]

1. Прочитаем данные из файла, сразу распределяя их по кластерам. При этом нужно учесть, что первая строка файла содержит заголовки (буквы X и Y), а в качестве разделителя целой и дробной части чисел использована запятая, которую нужно заменить на точку для нормальной работы функции float:

**with** open('27-A.txt') **as** F:

F.readline() # читаем и пропускаем заголовки

**for** s **in** F:

x, y = s.replace(',','.').split()

x, y = float(x), float(y)

clusterNo = findClusterNo( x, y )

clusters[clusterNo].append( (x, y) )

Теперь clusters[0] – это список пар (x, y) (кортежей), в которых записаны координаты всех точек кластера 0, а clusters[1] – такие же пары для точек кластера 1.

1. Следующий этап – в каждом кластере найти *центроид* – точку, для которой сумма расстояний до других точек кластера минимальна. Напишем функцию, которая вычисляет расстояние между точками, данные о которых передаются в виде двух кортежей:

**def** dist( p1, p2 ):

**return** ((p1[0] - p2[0])\*\*2 + (p1[1] - p2[1])\*\*2) \*\* 0.5

или так

**import** math

**def** dist( p1, p2 ):

**return** math.hypot( p1[0] - p2[0], p1[1] - p2[1] )

1. Создаём пустой список для хранения координат центров кластеров:

centers = []

1. Перебираем в цикле все кластеры, для каждого определяем центроид и добавляем его координаты в список centers:

**for** k **in** range(K):

...

centers.append( center )

1. В теле цикла перебираем все точки кластера; для каждой из них находим сумму расстояний до остальных точек sumDist и выбираем такую точку center, для которой это расстояние наименьшее

minSumDist = float('inf')

**for** pCenter **in** clusters[k]:

sumDist = sum( dist(pCenter,p)

**for** p **in** clusters[k] )

**if** sumDist < minSumDist:

minSumDist = sumDist

center = pCenter

1. Остаётся вычислить среднее значение координат центроидов (по каждой оси), умножить их на 10000 и найти целые части этих чисел:

sumX, sumY = 0, 0

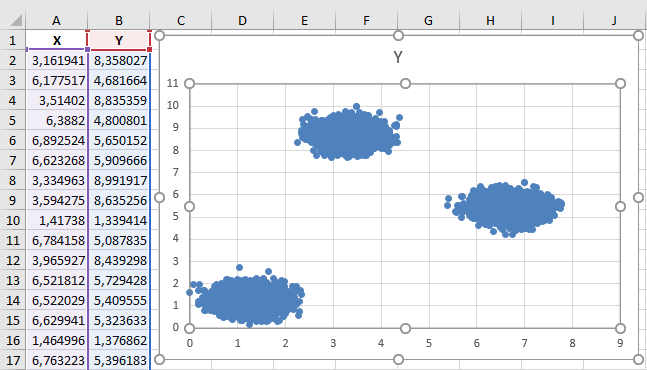
**fo**r k **in** range(K):

sumX += centers[k][0]

sumY += centers[k][1]

print( int(abs(sumX)/K\*10000), int(abs(sumY)/K\*10000) )

1. Для решения задачи Б построим точечную диаграмму по данным второго файла:



1. Единственная сложность – написать условия, разделяющие эти кластеры. Заметим, что самый левый кластер отделяется от двух других прямой *y* = 3, а два правых разделяются прямой *x* = 5. Таким образом, в программе нужно только изменить значение K и функцию findClusterNo:

K = 3 # количество кластеров

**def** findClusterNo( x, y ):

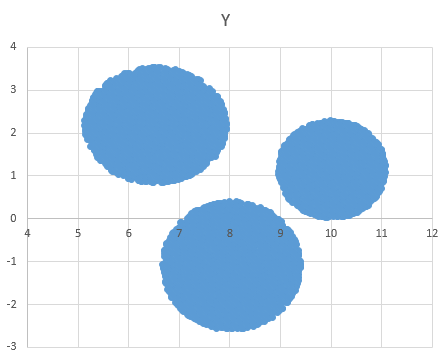
**return** 0 **if** y < 3 **else** \

1 **if** x < 5 **else** \

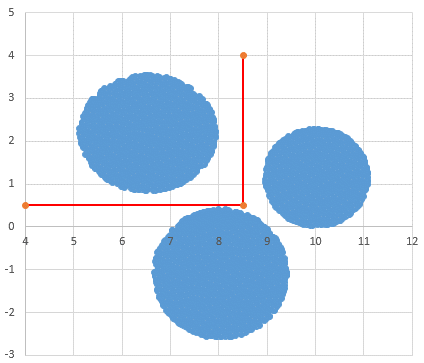
2

Остальные части программы не изменяются.

1. К сожалению, не всегда удаётся разделить кластеры, используя только одну координату (т. е. только горизонтальные и вертикальные прямые). Предположим, что распределение точек на плоскости выглядит примерно так:



Здесь легче всего сначала отделить кластер, который находится на диаграмме слева вверху – это можно сделать прямыми *x* = 8,5 и *y* = 0,5.



1. Затем разделяем оставшиеся два кластера прямой *y* = –*x* + 9,3 (отрезок этой прямой показан зелёным цветом):



Коэффициенты уравнения прямой подбираются методом проб и ошибок в табличном процессоре.

1. В результате функция, определяющая номер кластера для точки с координатами (*x*, *y*) выглядит так:

**def** findClusterNo( x, y ):

**return** 0 **if** y > 0.5 **and** x < 8.5 **else** \

1 **if** y > -x+9.3 **else** \

2

**Задача 2.** Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба. Кластер звёзд – это набор звёзд (точек) на графике, лежащий внутри круга радиусом . Каждая звезда обязательно принадлежит только одному из кластеров.

Истинный центр кластера, или центроид, – это одна из звёзд на графике, сумма расстояний от которой до всех остальных звёзд кластера минимальна. Центроид не вычисляется для колец, он вычисляется только для кластеров, представляющих собой круг.

Под расстоянием понимается расстояние Евклида между двумя точками  А(х1 ,у1) и В(х2 ,у2)  на плоскости, которое вычисляется по формуле:

В файле A хранятся данные о звёздах **двух**кластеров, где  для внутреннего кластера и  для внешнего кластера. В каждой строке записана информация о расположении на карте одной звезды: сначала координата , затем координата . Значения даны в условных единицах, которые представлены вещественными числами. Известно, что количество звёзд не превышает 1245.

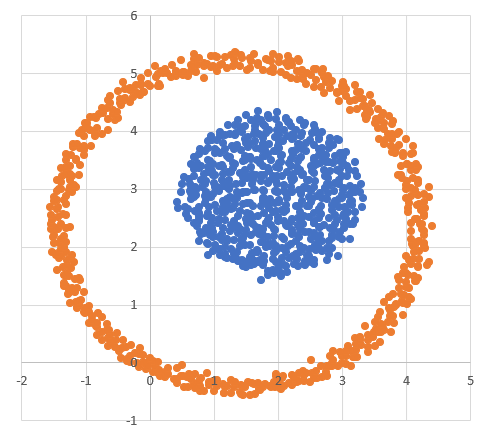
В файле Б хранятся данные о звёздах **трёх**кластеров, где  для двух внутренних кластеров и  для внешнего кластера. Известно, что количество звёзд не превышает 9414. Структура хранения информации о звездах в файле Б аналогична файлу А.

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа:  Px — среднее арифметическое абсцисс центров кластеров, и  Py – среднее арифметическое ординат центров кластеров.

В ответе запишите четыре числа через пробел: сначала целую часть произведения  Px⋅100 для файла А и  Py⋅100 для файла А, далее целую часть произведения  Px⋅100 для файла Б и  Py⋅100 для файла Б.

Возможные данные одного из файлов иллюстрированы графиком.

**Внимание! График приведён в иллюстративных целях для произвольных значений, не имеющих отношения к заданию. Для выполнения задания используйте данные из прилагаемого файла.**



**Вложения к задаче:**

[27\_1\_A.txt](https://3.shkolkovo.online/st/7/o/27_1_A__51u7z.txt" \t "_blank)

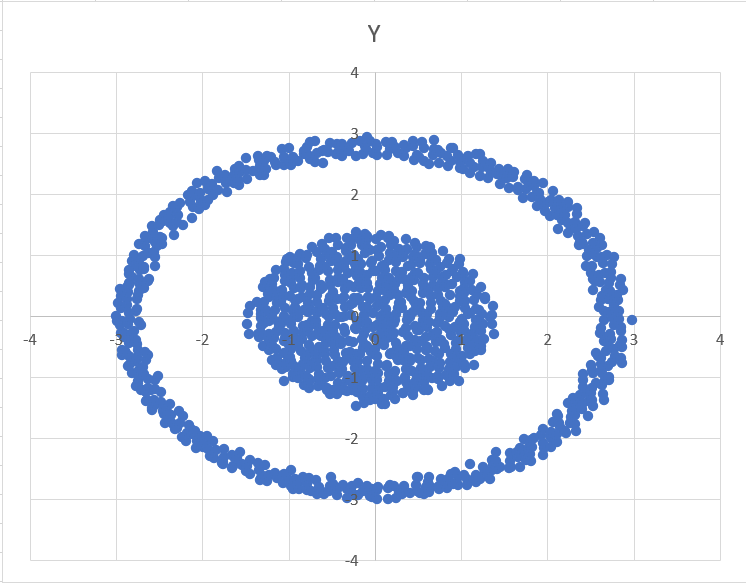
[27\_1\_A.xlsx](https://3.shkolkovo.online/st/7/o/27_1_A__51u80.xlsx" \t "_blank)

[27\_1\_B.txt](https://3.shkolkovo.online/st/7/o/27_1_B__51u81.txt" \t "_blank)

[27\_1\_B.xlsx](https://3.shkolkovo.online/st/7/o/27_1_B__51u82.xlsx" \t "_blank)

**Решение:** Для начала визуально оценим данные в условии кластеры. Для этого откроем предложенные файлы в электронной таблице, перейдем в раздел **«Вставка**  – **Диаграммы** – **Точечная»**.

Диаграмма для файла А имеет вид:



Внутренний кластер симметричен относительно оси Ox и Oy. Радиус внутреннего кластера равен 1.5, соответственно, та или иная звезда будет принадлежать данному кластеру если она удовлетворяет следующему неравенству окружности:

**Код программы для файла А:**

f = open(’27\_1\_A.txt’)

n = f.readline()

cluster = []  # Создаём список для кластера

for i in range(1245):  # Считываем звёзды из файла

    star = list(map(float, f.readline().replace(’,’, ’.’).split()))

    if star[0] \*\* 2 + star[1] \*\* 2 <= 1.5\*\*2:  # Звезда относится к внутреннему кластеру

        cluster.append(star)

tx = ty = 0

mn = 100000050000

for j in cluster:  # Перебор предполагаемого центроида

    x1, y1 = j

    sm = 0  # Суммарное расстояние

    for k in cluster:  # Перебор всех звёзд для вычисления суммарного расстояния

        x2, y2 = k

        sm += ((x2 - x1) \*\* 2 + (y2 - y1) \*\* 2) \*\* 0.5

    if sm < mn: # У текущей звезды расстояние до остальных звёзд меньше предыдущей

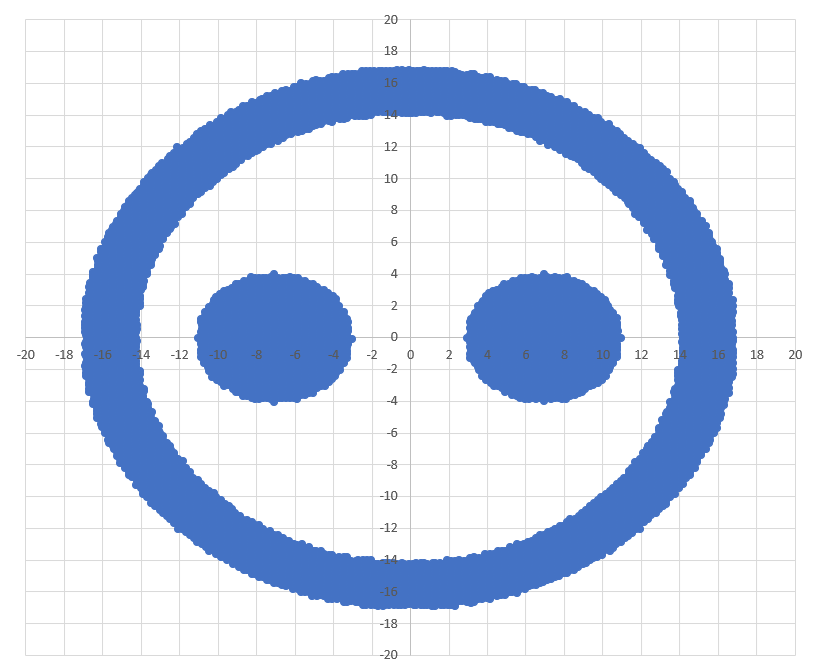
        mn = sm

        tx, ty = x1, y1 # Сохранение нового центроида

print(int(tx \* 100))

print(int(ty \* 100))

Диаграмма для файла Б имеет вид:



Определим подходящие уравнения окружностей, внутри которых лежат внутренние кластеры. Геометрические центры левого и правого кластеров находятся около точки (-7; 0) и (7; 0) соответственно, и для них можно выбрать радиус 4.5, не задевая внешний кластер. Получим следующие неравенства для внутренних кластеров:

 – для левого кластера

 – для правого кластера

**Код программы для файла Б:**

f = open(’27\_1\_B.txt’)

n = f.readline()

clusters = [[] for i in range(2)]

for i in range(9414):

    star = list(map(float, f.readline().replace(’,’, ’.’).split()))

    if (star[0] - 7) \*\* 2 + star[1] \*\* 2 <= 4.5 \*\* 2:

        clusters[0].append(star)

    elif (star[0] + 7) \*\* 2 + star[1] \*\* 2 <= 4.5 \*\* 2:

        clusters[1].append(star)

sum\_x = sum\_y = 0

for i in clusters:

    tx = ty = 0

    mn = 100000050000

    for j in i:

        x1, y1 = j

        sm = 0

        for k in i:

            x2, y2 = k

            sm += ((x2 - x1) \*\* 2 + (y2 - y1) \*\* 2) \*\* 0.5

        if sm < mn:

            mn = sm

            tx, ty = x1, y1

    sum\_x += tx

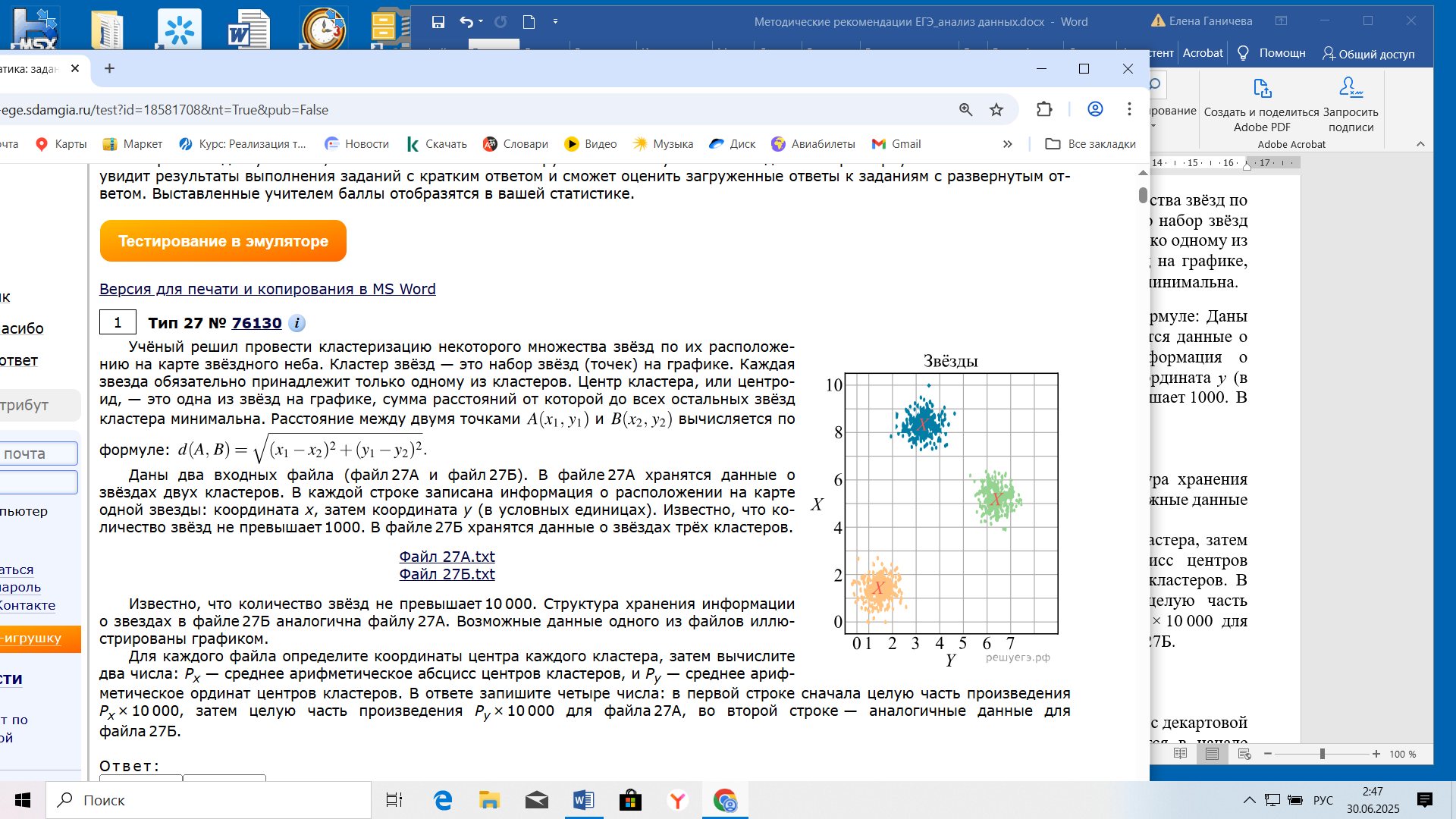
    sum\_y += ty

print(int(sum\_x / len(clusters) \* 100))

print(int(sum\_y / len(clusters) \* 100))

**Ответ: -8 -9 -5 -4**

1. **Задачи для самостоятельного решения**
2. Учёный решил провести кластеризацию некоторого множества звёзд по их расположению на карте звёздного неба. Кластер звёзд  — это набор звёзд (точек) на графике. Каждая звезда обязательно принадлежит только одному из кластеров. Центр кластера, или центроид,  — это одна из звёзд на графике, сумма расстояний от которой до всех остальных звёзд кластера минимальна.

Расстояние между двумя точками

А(х1 ,у1) и В(х2 ,у2) вычисляется по формуле:

Даны два входных файла (файл 27A и файл 27Б). В файле 27A хранятся данные о звёздах двух кластеров. В каждой строке записана информация о расположении на карте одной звезды: координата *x*, затем координата *y* (в условных единицах). Известно, что количество звёзд не превышает 1000. В файле 27Б хранятся данные о звёздах трёх кластеров.

[Файл 27А.txt](https://inf-ege.sdamgia.ru/get_file?id=203410)

[Файл 27Б.txt](https://inf-ege.sdamgia.ru/get_file?id=203411)

Известно, что количество звёзд не превышает 10 000. Структура хранения информации о звездах в файле 27Б аналогична файлу 27А. Возможные данные одного из файлов иллюстрированы графиком.

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: *Px*  — среднее арифметическое абсцисс центров кластеров, и *Py*  — среднее арифметическое ординат центров кластеров. В ответе запишите четыре числа: в первой строке сначала целую часть произведения |*Px* |× 10 000, затем целую часть произведения |*Py* |× 10 000 для файла 27А, во второй строке  — аналогичные данные для файла 27Б.

Ответ для A: 32055 58097.

Ответ для B: 31886 25834.

1. Фрагмент звёздного неба спроецирован на плоскость с декартовой системой координат. Учёный решил провести кластеризацию полученных точек, являющихся изображениями звёзд, то есть разбить их множество на *N* непересекающихся непустых подмножеств (кластеров), таких, что точки каждого подмножества лежат внутри квадрата со стороной длиной *H*, причём эти квадраты между собой не пересекаются. Стороны квадрата не обязательно параллельны координатным осям. Гарантируется, что такое разбиение существует и единственно для заданных размеров квадрата.

Будем называть центром кластера точку этого кластера, сумма расстояний от которой до всех остальных точек кластера минимальна. Для каждого кластера гарантируется единственность его центра. Расстояние между двумя точками на плоскости А(х1 ,у1) и В(х2 ,у2) вычисляется по формуле:

В файле А хранятся координаты точек **двух** кластеров, где *H*  =  4,7 для каждого кластера. В каждой строке записана информация о расположении на карте одной звезды: сначала координата *x*, затем координата *y*. Известно, что количество точек не превышает 1000.

В файле Б хранятся координаты точек **трёх** кластеров, где *H*  =  4 для каждого кластера. Известно, что количество точек не превышает 10 000. Структура хранения информации в файле Б аналогична файлу A.

[Файл A](https://inf-ege.sdamgia.ru/get_file?id=202990)

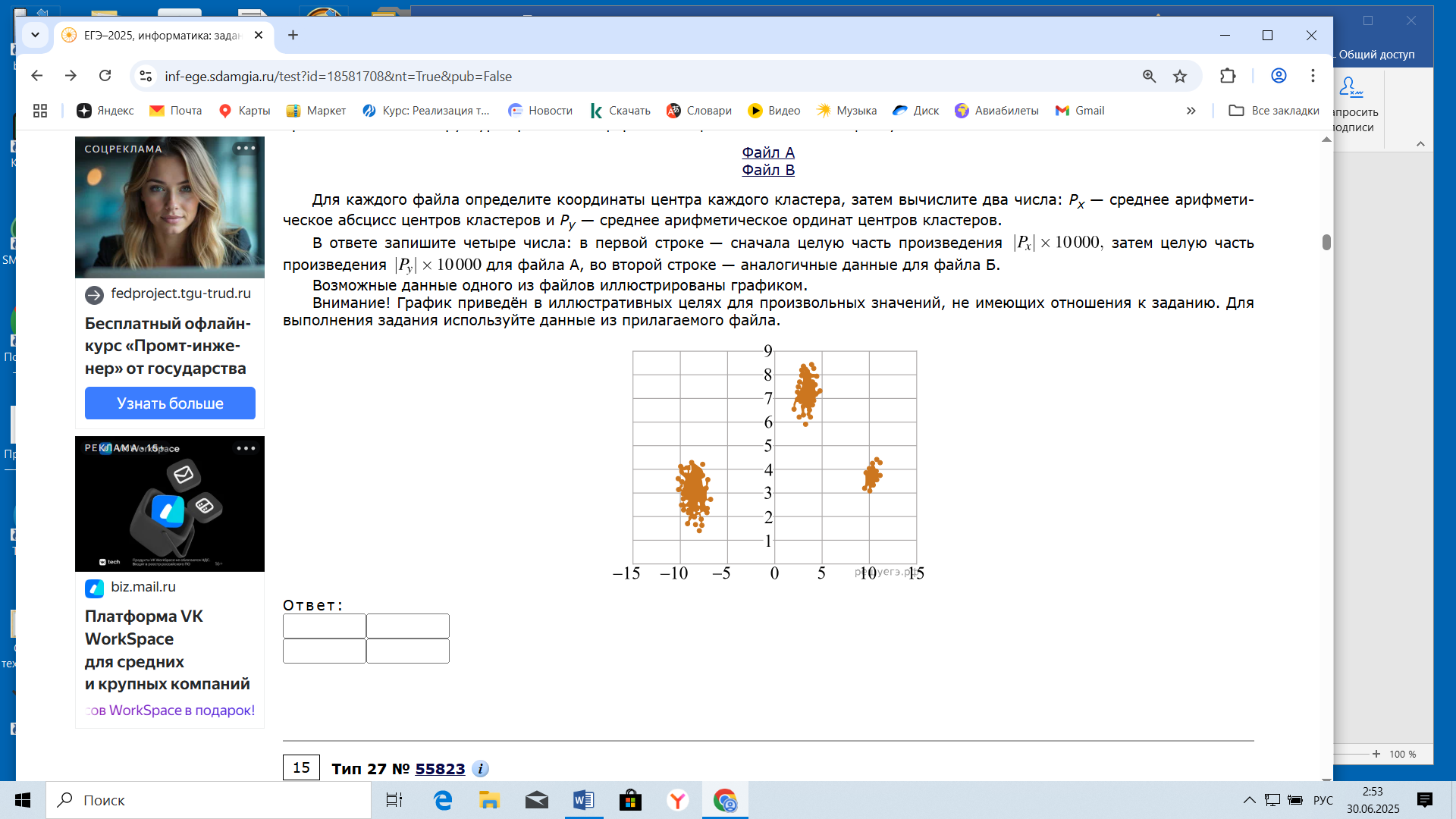
[Файл B](https://inf-ege.sdamgia.ru/get_file?id=202991)

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: *Px*  — среднее арифметическое абсцисс центров кластеров и *Py*  — среднее арифметическое ординат центров кластеров.

В ответе запишите четыре числа: в первой строке  — сначала целую часть произведения |*Px* |× 10 000, затем целую часть произведения |*Py* |× 10 000 для файла A, во второй строке  — аналогичные данные для файла Б.

Возможные данные одного из файлов иллюстрированы графиком.

Внимание! График приведён в иллюстративных целях для произвольных значений, не имеющих отношения к заданию. Для выполнения задания используйте данные из прилагаемого файла.



Ответ для A: 43598 166165.

Ответ для B: 17382 30732.

1. Фрагмент звёздного неба спроецирован на плоскость с декартовой системой координат. Учёный решил провести кластеризацию полученных точек, являющихся изображениями звёзд, то есть разбить их множество на *N* непересекающихся непустых подмножеств (кластеров), таких, что точки каждого подмножества лежат внутри квадрата со стороной длиной *H*, причём эти квадраты между собой не пересекаются. Стороны квадрата не обязательно параллельны координатным осям. Гарантируется, что такое разбиение существует и единственно для заданных размеров квадрата.

Будем называть центром кластера точку этого кластера, сумма расстояний от которой до всех остальных точек кластера минимальна. Для каждого кластера гарантируется единственность его центра. Расстояние между двумя точками на плоскости А(х1 ,у1) и В(х2 ,у2) вычисляется по формуле:

В файле А хранятся координаты точек **двух** кластеров, где *H*  =  4,7 для каждого кластера. В каждой строке записана информация о расположении на карте одной звезды: сначала координата *x*, затем координата *y*. Известно, что количество точек не превышает 1000.

В файле Б хранятся координаты точек **трёх** кластеров, где *H*  =  4 для каждого кластера. Известно, что количество точек не превышает 10 000. Структура хранения информации в файле Б аналогична файлу A.

[Файл A](https://inf-ege.sdamgia.ru/get_file?id=202990)

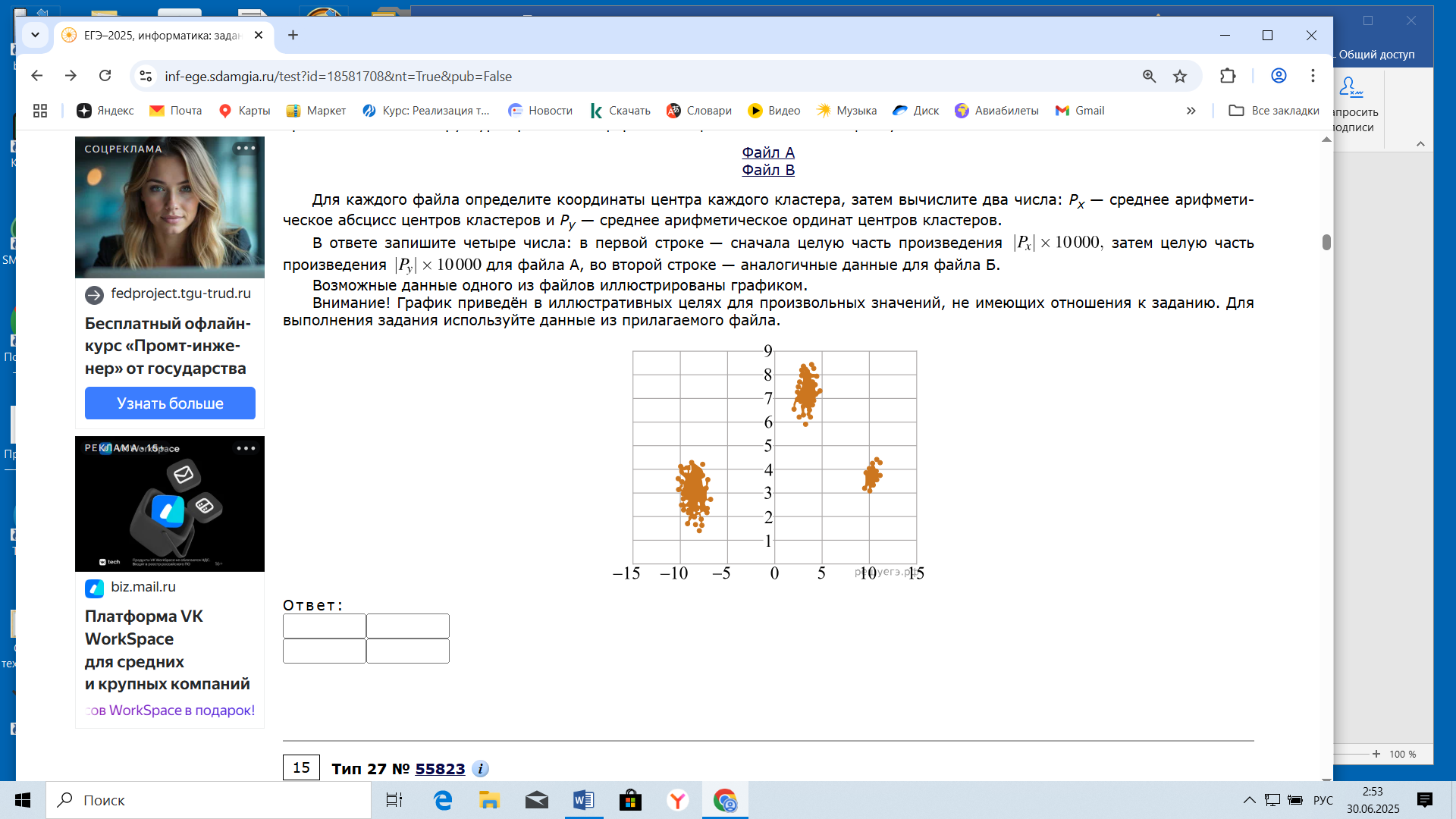
[Файл B](https://inf-ege.sdamgia.ru/get_file?id=202991)

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: *Px*  — среднее арифметическое абсцисс центров кластеров и *Py*  — среднее арифметическое ординат центров кластеров.

В ответе запишите четыре числа: в первой строке  — сначала целую часть произведения |*Px* |× 10 000, затем целую часть произведения |*Py* |× 10 000 для файла A, во второй строке  — аналогичные данные для файла Б.

Возможные данные одного из файлов иллюстрированы графиком.

Внимание! График приведён в иллюстративных целях для произвольных значений, не имеющих отношения к заданию. Для выполнения задания используйте данные из прилагаемого файла.



Ответ для A: 56173 138075.

Ответ для B: 28283 33540.

1. Фрагмент звёздного неба спроецирован на плоскость с декартовой системой координат. Учёный решил провести кластеризацию полученных точек, являющихся изображениями звёзд, то есть разбить их множество на *N* непересекающихся непустых подмножеств (кластеров), таких, что точки каждого подмножества лежат внутри квадрата со стороной длиной *H*, причём эти квадраты между собой не пересекаются. Стороны квадрата не обязательно параллельны координатным осям. Гарантируется, что такое разбиение существует и единственно для заданных размеров квадрата.

Будем называть центром кластера точку этого кластера, сумма расстояний от которой до всех остальных точек кластера минимальна. Для каждого кластера гарантируется единственность его центра. Расстояние между двумя точками на плоскости А(х1 ,у1) и В(х2 ,у2) вычисляется по формуле:

В файле А хранятся координаты точек **двух** кластеров, где *H*  =  4,7 для каждого кластера. В каждой строке записана информация о расположении на карте одной звезды: сначала координата *x*, затем координата *y*. Известно, что количество точек не превышает 1000.

В файле Б хранятся координаты точек **трёх** кластеров, где *H*  =  4 для каждого кластера. Известно, что количество точек не превышает 10 000. Структура хранения информации в файле Б аналогична файлу A.

[Файл A](https://inf-ege.sdamgia.ru/get_file?id=202990)

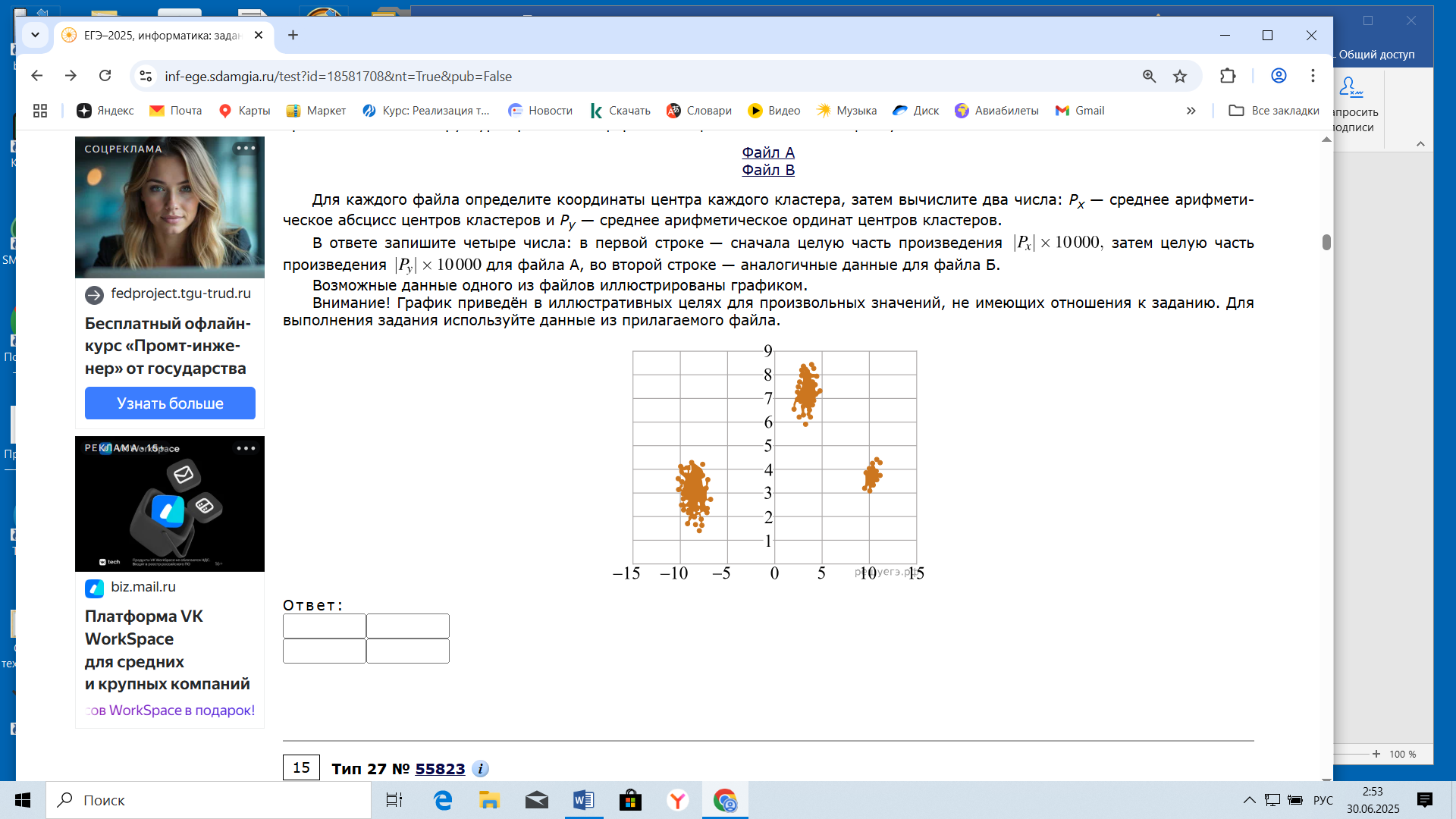
[Файл B](https://inf-ege.sdamgia.ru/get_file?id=202991)

Для каждого файла определите координаты центра каждого кластера, затем вычислите два числа: *Px*  — среднее арифметическое абсцисс центров кластеров и *Py*  — среднее арифметическое ординат центров кластеров.

В ответе запишите четыре числа: в первой строке  — сначала целую часть произведения |*Px* |× 10 000, затем целую часть произведения |*Py* |× 10 000 для файла A, во второй строке  — аналогичные данные для файла Б.

Возможные данные одного из файлов иллюстрированы графиком.

Внимание! График приведён в иллюстративных целях для произвольных значений, не имеющих отношения к заданию. Для выполнения задания используйте данные из прилагаемого файла.



Ответ для A: 54425 133945.

Ответ для B: 30768 35102.

**Заключение**

Подготовка к решению задач анализа данных позволяет сформировать навыки работы с инструментами электронной таблицы, такие, как умение построить диаграмму, выполнить сортировку данных, разные способы вставки данных, умение записать формулу, умение применять встроенные функции, условное форматирование [6,7,10,11]. Уверенное владение инструментами электронной таблицы позволит выпускнику успешно применять их как для решения задач в практической деятельности, так и при выполнении заданий итоговой государственной аттестации [12].

Следует отметить, что освоение методических приемов работы над заданием №27 ЕГЭ по информатике развивает способность логически мыслить, анализировать и интерпретировать данные, создавать математические модели решения практических задач.

В методическом кейсе представлены способы решения ключевых задач, предлагается подборка задач для самостоятельного решения на закрепление полученных умений.

**Список литературы:**

1. Богатырева Ю. И. Направления решения проблемы недостаточной подготовки школьников к ЕГЭ по информатике в Тульской области / Ю. И. Богатырева, К. Г. Гончаров, О. В. Родионова // Педагогическая информатика. 2023. № 4. С. 38-48. EDN: HONUDZ.
2. Везиров Т. Т. Средства электронного обучения в подго товке к ЕГЭ по информатике и ИКТ // Педагогический журнал. 2018. Т. 8. № 6A. С. 225–232. EDN: ZAHILB.
3. Ганичева Е. М., Голубев О.Б. Работа над задачами по теме "Использование динамических (электронных) таблиц для выполнения учебных заданий" при подготовке обучающихся к ГИА по информатике : учебное пособие для подготовки к государственной итоговой аттестации выпускников основной и старшей школы. Вологда: Вологодский институт развития образования, 2022. 76 с. EDN: WHAOJQ.
4. Демоверсии, спецификации, кодификаторы ЕГЭ 2025 год. <https://fipi.ru/ege/demoversii-specifikacii-kodifikatory> .
5. Зубрилин А. А., Хохлова А. Д. Возможности табличного процессора MS Excel при решении задач ЕГЭ по информатике (на примере задачи № 5) // Инновационное образовательное пространство педагогического вуза: проблемы и перспективы. Сборник статей по материалам Всероссийской научно-практической конференции. Саранск: Мордовский государственный педагогический университет им. М. Е. Евсевьева, 2023. С. 80–86. EDN: ANLAKS.
6. Зубрилин А.А. Возможности табличных процессоров при подготовке к ЕГЭ по информатике // Информатика в школе. 2024. № 6. С. 78–86. EDN: GOOTWX. DOI: 10.32517/2221-1993-2024-6--78-86.
7. Ильина Е. А., Засов В. О., Масальский Л. С., Логунова Т. В., Шишиморов А. П. О генерации визуальных материалов к ЕГЭ по информатике // Педагогическая информатика. 2023. № 2. С. 25–28. EDN: KMEJPH.
8. Касторнова, В. А. К вопросу о внедрении технологий искусственного интеллекта в школьное образование // Педагогическая информатика. 2022. № 1. С. 18-29. EDN: NDWHAN.
9. Чикова О. А., Каменев Р. В., Витюнин М. А., Сартако И. В. Методика решения прикладных задач на уроках информатики в профильных информационно-технологических классах // Педагогическая информатика. 2023. № 4. С. 58-75. EDN: EOEZEC.
10. Сдам ГИА: Решу ЕГЭ. Образовательный портал для подготовки к экзаменам. Информатика. <https://inf-ege.sdamgia.ru> .
11. Табачук Н. П., Бусарова У. А. Образовательный контент ЕГЭ по информатике в картах визуализации // Педагогическая информатика. 2023. № 2. С. 63–71. EDN: RFORQK.
12. Фирсова С. А. Особенности решения некоторых задач компьютерного ЕГЭ по информатике // Вестник Таганрогского института имени А. П. Чехова. 2021. № 2. С. 47–51. EDN: RMJWOD.

**Источники:**

1. Открытый банк заданий ЕГЭ https://ege.fipi.ru/bank/
2. **Методические рекомендации для учителей, подготовленные на основе анализа типичных ошибок участников ЕГЭ 2023 года** https://doc.fipi.ru/ege/analiticheskie-i-metodicheskie-materialy/2023/inf\_mr\_2023.pdf
3. Демоверсия, спецификация, кодификатор ЕГЭ 2024

https://fipi.ru/ege/demoversii-specifikacii-kodifikatory#!/tab/151883967-5

1. Сайт Сдам ГИА/Решу ЕГЭ <https://inf-ege.sdamgia.ru/>
2. Блог по информатике Code-enjoy https://code-enjoy.ru/
3. Всероссийский проект «ЕГЭ 100 БАЛЛОВ» <https://vk.com/ege100ballov>