Задание 1

По определению первым коэффициентом вейвлета является выборочное среднее:

$$\psi_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = 2.9375 \quad (1)$$

Далее вычисляются средние двух половин выборки и в качестве второго коэффициента вейвлета Хаара берется полуразность средних левой и правой половин:

$$\psi_2 = \frac{1}{2}\left(\overline{x}_{left} - \overline{x}_{right}\right) = \frac{1}{2}(4.375 - 1.5) = 1.4375 \quad (2)$$

Далее алгоритм продолжается рекурсивно для половин выборки. То есть для левой половины выборки вычисляется полуразность уже ее правой и левой половины, равная $\psi_3 = -7.375$. Аналогично поступаем для правой половины, получая 4-й коэффициент: $\psi_4 = 8.25$.

Вычисленные коэффициенты собраны в таблице 1, где сверху вниз в левом столбце вычисляются средние половин выборки, а в правом столбце приведены соответствующие коэффициенты вейвлета Хаара.

Таблица 1. Средние значения выборки и соответствующие коэффициенты вейвлета Хаара

Средние значения выборки	Коэф-ты Вейвлета
$\{-25, -11, -8, -3, 1, 0, 3, 8, 11, 13, 9, 6, 2, -2, -9, -18\}$	$-7, -2.5, 0.5, -2.5, -1, -1.5, 2{,}5, 4.5$
$\{-18, -5.5, -0.5, 5.5, 12, 7.5, 0, -13.5\}$	$-6.25, -2.5, 2.25, 6.75$
$\{-11.75, 3, 9.75, -6.75\}$	$-7.375, 8.25$
$\{4.375, 1.5\}$	1.4375
$\{2.9375\}$	2.9375

Ответ: $2.9375, 1.4375, -7.375, 8.25, -6.25, -2.5, 2.25, 6.75, -7, -2.5, 0.5, -2.5, -1, -1.5, 2.5, 4.5$.

Задание 2

Таблица 2. Выборка

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
x_i	1,7	0,4	0,6	2,2	0,6	2,8	2,2	2,2	2,4

1. Вычислим оптимальное количество интервалов по правилу Стерджесса: $n\_bins = 1 + +\lfloor log_2 n \rfloor = 1 + \lfloor log_2 9 \rfloor = 4$. Гистограмма показана на рисунке 1.

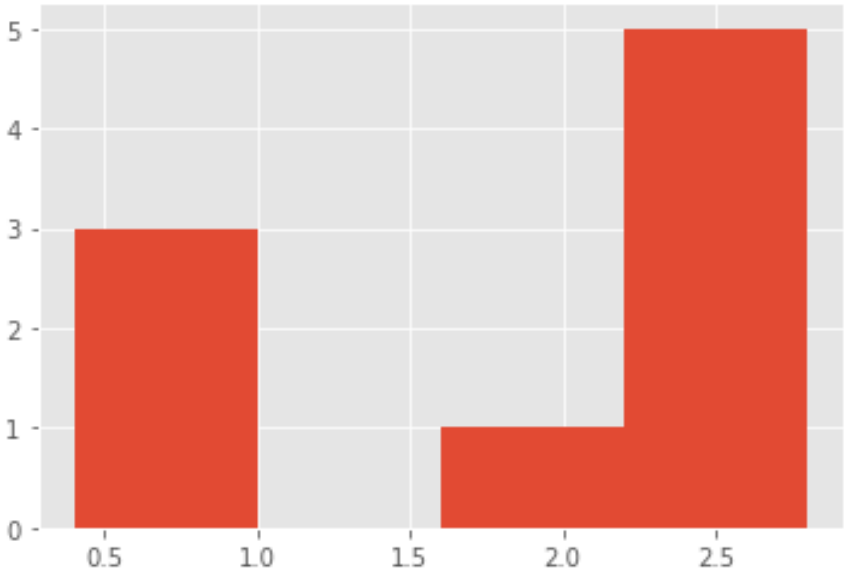

Рисунок 1 – гистограмма выборки

2. Модой выборки является элемент 2.2. Рассчитаем веса элементов выборки (пока не нормированные на 1) по формуле (значения приведены в таблице 3):

$$\widetilde{w}_i = \frac{1}{(x_i - m)^2} \tag{3}$$

Таблица 3. Невзвешенные коэффициенты.

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
x_i	1,7	0,4	0,6	2,2	0,6	2,8	2,2	2,2	2,4
$\widetilde{w}_i$	4	0.31	0.39	-	0.39	2.78	-	-	25

Значения в таблице округлены до сотых для наглядности, но при дальнейших расчетах точность сохранена. Прочерки на месте элементов, равных моде, обозначают, что их веса $(1/9)$ полагаются сразу нормированными. Условие нормировки в данном случае будет выглядеть следующим образом:

$$\frac{k}{n} + a \cdot \sum_{i \in I} \widetilde{w}_i = 1, I = \{i : x_i \neq m\} \tag{4}$$

где a – нормировочный множитель, m – мода выборки, $\widetilde{w}_i$ – ненормированный вес элемента x_i, k – количество элементов выборки, равных моде.

Рассчитаем нормировочный коэффициент из условия нормировки (2): $a = 0.0202833566138063$.

Теперь рассчитаем нормированные на 1 веса, умножив ненормированные на нормировочный коэффициент (в таблице 4 округлены до тысячных).

Таблица 4. Взвешенные коэффициенты

i	1	2	3	4	5	6	7	8	9
x_i	1,7	0,4	0,6	2,2	0,6	2,8	2,2	2,2	2,4
w_i	0.081	0.006	0.008	0.111	0.008	0.056	0.111	0.111	0.507

Теперь рассчитаем взвешенное среднее (надо по-хорошему еще на сумму весов делить, но она у нас равна 1, поэтому опускаю это):

$$\overline{x}_w = \sum_{i=1}^{n} w_i \cdot x_i = 2.25803293697839 \quad (5)$$

3. Рассчитаем выборочную дисперсию по формуле:

$$\widehat{D} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad (6)$$

где $\overline{x}$ – выборочная оценка математического ожидания (то есть среднее арифметическое). Стоит заметить, что данная оценка дисперсии является смещенной (для несмещенной оценки требуется делить не на n, а на $n-1$). Несложными вычислениями путем подстановки чисел получаем, что $\widehat{D} = 0.728395061728395$.

Выборочное среднеквадратичное отклонение является корнем из выборочной дисперсии, т.е. $\widehat{\sigma} = \sqrt{\widehat{D}} \approx 0.85346$.

Ответ: 1) См. рисунок 1; 2) $\overline{x}_w = 2.25803293697839$; 3) $\widehat{D} = 0.728395061728395$, $\hat{\sigma} \approx 0.85346$.

Задание 3

Минимальный уровень поддержки: 30%. Учитывая, что всего транзакций 10, то это обозначает, что часто встречающимися будут являться такие подмножества, которые присутствуют в 3 и более транзакциях.

Таблица 5. Количество транзакций, включающих 1-наборы $\{I_j\}$

Подмн-во	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	I16
Кол-во транзакций	4	2	4	1	3	2	1	4	3	4	4	2	2	1	4	3

Получается, что множество $C_1 = L_1 = \{\{I_1\},\{I_3\},\{I_5\},\{I_8\},\{I_9\},\{I_{10}\},\{I_{11}\},\{I_{15}\},\{I_{16}\}\}$.

Далее вычисляем мн-во кандидатов 2-наборов: $C_2 = L_1 \rhd\lhd L_1$.

Таблица 6. Количество транзакций, включающих 2-наборы $\{I_j, I_k\}, j < k$

Подмн-во	{I1,I3}	{I1,I5}	{I1,I8}	{I1,I9}	{I1,I10}	{I1,I11}	{I1,I15}	{I1,I16}	{I3,I5}
Кол-во транзакций	0	0	0	0	2	2	3	2	3

Подмн-во	{I3,I8}	{I3,I9}	{I3,I10}	{I3,I11}	{I3,I15}	{I3,I16}	{I5,I8}	{I5,I9}	{I5,I10}
Кол-во транзакций	4	3	0	0	0	1	3	2	0

Подмн-во	{I5,I11}	{I5,I15}	{I5,I16}	{I8,I9}	{I8,I10}	{I8,I11}	{I8,I15}	{I8,I16}
Кол-во транзакций	0	0	1	3	0	0	0	1

Подмн-во	{I9,I10}	{I9,I11}	{I9,I15}	{I9,I16}	{I10,I11}	{I10,I15}	{I10,I16}	{I11,I15}
Кол-во транзакций	0	0	0	1	4	3	0	3

Подмн-во	{I11,I16}	{I15,I16}
Кол-во транзакций	0	1

Множество выбранных по мин. уровню поддержки получается таким: $L_2 = \{\{I_1, I_{15}\}, \{I_3, I_8\}, \{I_3, I_9\}, \{I_5, I_8\}, \{I_8, I_9\}, \{I_{10}, I_{11}\}, \{I_{10}, I_{15}\}, \{I_{11}, I_{15}\}\}$.

Рассчитаем мн-во кандидатов на частые 3-наборы $C_3 = L_2 \rhd\lhd L_2$.

Таблица 7. Количество транзакций, включающих 3-наборы $\{I_j, I_k, I_l\}, j < k < l$

Подмножество	{I3,I8,I9}	{I10,I11,I15}
Кол-во транзакций	3	3

Множество часто встречающихся 3-наборов: $L_3 = \{\{I_3, I_8, I_9\}, \{I_{10}, I_{11}, I_{15}\}\}$.

Произвести слияние L_3 с собой не получится, так как $\{I_3, I_8\} \neq \{I_{10}, I_{11}\}$.

Получается, что часто встречающиеся наборы ограничиваются следующим множеством: $L = \bigcup_k L_k = \{\{I_1\}, \{I_3\}, \{I_5\}, \{I_8\}, \{I_9\}, \{I_{10}\}, \{I_{11}\}, \{I_{15}\}, \{I_{16}\}, \{I_1, I_{15}\}, \{I_3, I_8\}, \{I_3, I_9\}, \{I_5, I_8\}, \{I_8, I_9\}, \{I_{10}, I_{11}\}, \{I_{10}, I_{15}\}, \{I_{11}, I_{15}\}, \{I_3, I_8, I_9\}, \{I_{10}, I_{11}, I_{15}\}\}$

Ответ: $L = \{\{I_1\}, \{I_3\}, \{I_5\}, \{I_8\}, \{I_9\}, \{I_{10}\}, \{I_{11}\}, \{I_{15}\}, \{I_{16}\}, \{I_1, I_{15}\}, \{I_3, I_8\}, \{I_3, I_9\}, \{I_5, I_8\}, \{I_8, I_9\}, \{I_{10}, I_{11}\}, \{I_{10}, I_{15}\}, \{I_{11}, I_{15}\}, \{I_3, I_8, I_9\}, \{I_{10}, I_{11}, I_{15}\}\}$

Задание 4

Для каждого n-набора ($n > 1$, так как в 1-наборе невозможно выделить антецедент и консеквент) выделим все возможные антецеденты и консеквенты и посмотрим, превосходит ли уровень уверенности минимальный (35%). Если да, то правило принимается.

Стоит отметить, что превосходство уровня уверенности правила $A \Rightarrow B$ (где $A, B \subset C, B = C \backslash A$ для некоторого часто встречающегося набора C) над минимальным можно проверить по следующей формуле:

$$conf_{min} \leq conf(A \Rightarrow B) = \frac{support(\mathrm{C})}{support(A)} \quad (7)$$

В таблице 8 представлены возможные правила и уровень уверенности для них, рассчитанные по формуле (7).

Таблица 8. Уровень уверенности для ассоциативных правил

$A \Rightarrow B$	$I_1 \Rightarrow I_{15}$	$I_{15} \Rightarrow I_1$	$I_3 \Rightarrow I_8$	$I_8 \Rightarrow I_3$	$I_3 \Rightarrow I_9$	$I_9 \Rightarrow I_3$
$conf(A \Rightarrow B)$	0.75	0.75	1	1	0.75	1
$A \Rightarrow B$	$I_5 \Rightarrow I_8$	$I_8 \Rightarrow I_5$	$I_8 \Rightarrow I_9$	$I_9 \Rightarrow I_8$	$I_{10} \Rightarrow I_{11}$	$I_{11} \Rightarrow I_{10}$
$conf(A \Rightarrow B)$	1	0.75	0.75	1	1	1
$A \Rightarrow B$	$I_{10} \Rightarrow I_{15}$	$I_{15} \Rightarrow I_{10}$	$I_{11} \Rightarrow I_{15}$	$I_{15} \Rightarrow I_{11}$		
$conf(A \Rightarrow B)$	0.75	0.75	0.75	0.75		
$A \Rightarrow B$	$I_3 \Rightarrow \{I_8, I_9\}$	$\{I_8, I_9\} \Rightarrow I_3$	$I_8 \Rightarrow \{I_3, I_9\}$	$\{I_3, I_9\} \Rightarrow I_8$	$I_9 \Rightarrow \{I_3, I_8\}$	
$conf(A \Rightarrow B)$	0.75	1	0.75	1	1	
$A \Rightarrow B$	$\{I_3, I_8\} \Rightarrow I_9$	$I_{10} \Rightarrow \{I_{11}, I_{15}\}$	$\{I_{11}, I_{15}\} \Rightarrow I_{10}$	$I_{11} \Rightarrow \{I_{10}, I_{15}\}$	$\{I_{10}, I_{15}\} \Rightarrow I_{11}$	
$conf(A \Rightarrow B)$	0.75	0.75	1	0.75	1	
$A \Rightarrow B$	$I_{15} \Rightarrow \{I_{10}, I_{11}\}$		$\{I_{10}, I_{11}\} \Rightarrow I_{15}$			
$conf(A \Rightarrow B)$	0.75		0.75			

Ответ: $I_1 \Rightarrow I_{15}$, $I_{15} \Rightarrow I_1$, $I_3 \Rightarrow I_8$, $I_8 \Rightarrow I_3$, $I_3 \Rightarrow I_9$, $I_9 \Rightarrow I_3$, $I_5 \Rightarrow I_8$, $I_8 \Rightarrow I_5$, $I_8 \Rightarrow I_9$, $I_9 \Rightarrow I_8$, $I_{10} \Rightarrow I_{11}$, $I_{11} \Rightarrow I_{10}$, $I_{10} \Rightarrow I_{15}$, $I_{15} \Rightarrow I_{10}$, $I_{11} \Rightarrow I_{15}$, $I_{15} \Rightarrow I_{11}$, $I_3 \Rightarrow \{I_8, I_9\}$, $\{I_8, I_9\} \Rightarrow I_3$, $I_8 \Rightarrow \{I_3, I_9\}$, $\{I_3, I_9\} \Rightarrow I_8$, $I_9 \Rightarrow \{I_3, I_8\}$, $\{I_3, I_8\} \Rightarrow I_9$, $I_{10} \Rightarrow \{I_{11}, I_{15}\}$, $\{I_{11}, I_{15}\} \Rightarrow I_{10}$, $I_{11} \Rightarrow \{I_{10}, I_{15}\}$, $\{I_{10}, I_{15}\} \Rightarrow I_{11}$, $I_{15} \Rightarrow \{I_{10}, I_{11}\}$, $\{I_{10}, I_{11}\} \Rightarrow I_{15}$.

Задание 5

В таблице 9 представлены уровни поддержки для каждого 1-набора.

Таблица 9. Уровни поддержки 1-наборов

Подмн-во	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	I16
Кол-во транзакций	3	3	2	2	1	2	1	3	5	3	5	4	3	4	4	6

Теперь отсортируем их в порядке убывания поддержки: $\{I_{16}: 6\}, \{I_9: 5\}, \{I_{11}: 5\}, \{I_{12}: 4\}, \{I_{14}: 4\}, \{I_{15}: 4\}, \{I_1: 3\}, \{I_2: 3\}, \{I_8: 3\}, \{I_{10}: 3\}, \{I_{13}: 3\}, \{I_3: 2\}, \{I_4: 2\}, \{I_6: 2\}$

Отсортируем элементы транзакций в порядке убывания поддержки (таблица 10) после чего построим FP-дерево (рис. 2), при этом выкидываем те элементы, которые не удовлетворяют минимальному уровню поддержки.

Таблица 10. Отсортированные транзакции

Номер транзакции	Отсортированные элементы
1	I16, I9, I11, I15, I4
2	I16, I1, I2
3	I16, I9, I14, I1, I2, I10, I13, I3
4	I11, I12, I15, I6
5	I16, I9, I14, I10, I13, I4
6	I12, I15, I8
7	I9, I11, I8
8	I11, I12, I14, I15, I3, I6
9	I16, I1, I10, I13
10	I16, I9, I11, I12, I14, I2, I8

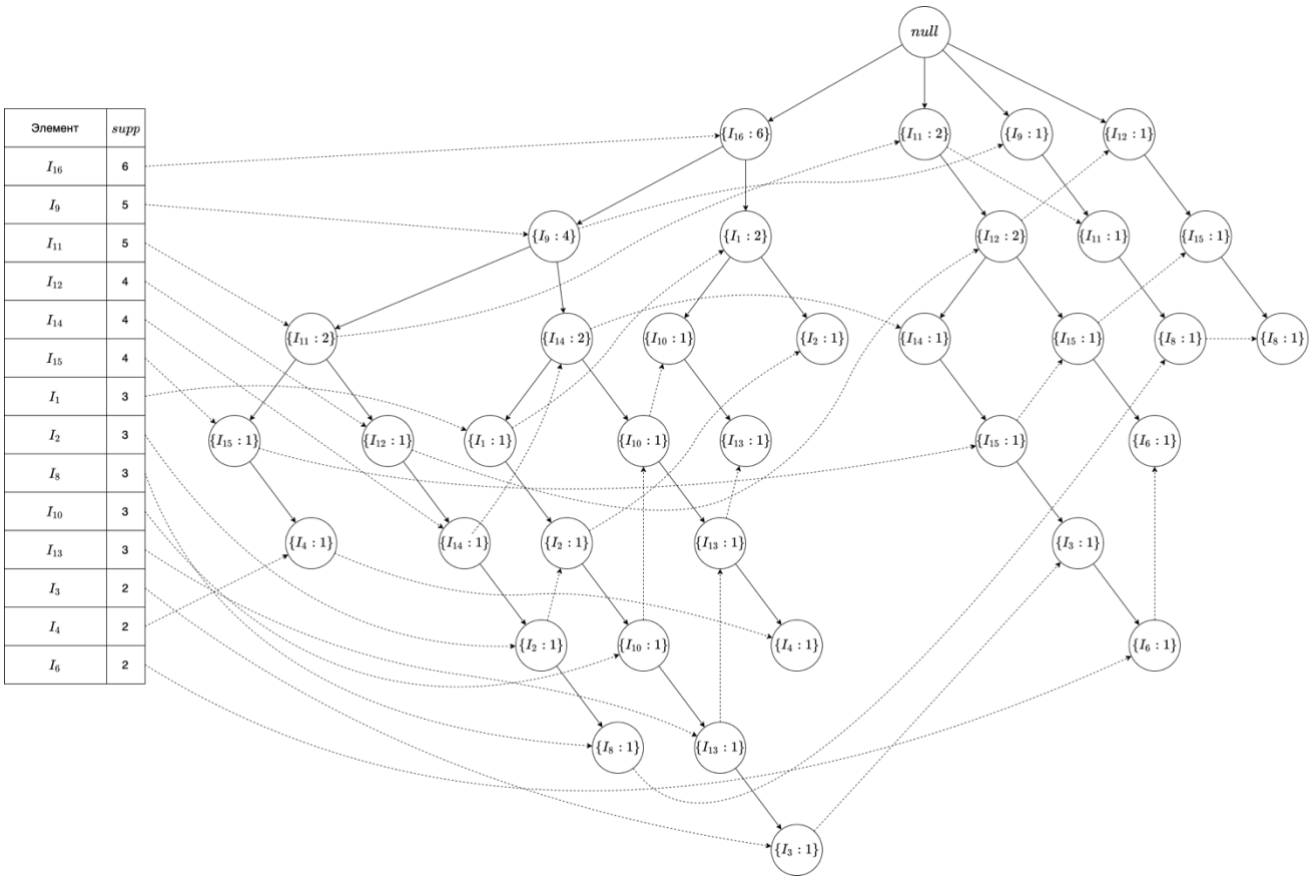

Рисунок 2 – FP-дерево

Теперь будем рассматривать элементы и строить условные FP-деревья для определения часто встречающихся подмножеств (табл. 11). Условное FP-дерево строится как и обычное выше,

но с условной базой шаблонов в качестве базы транзакций. Рассмотрим построение для I_4 (для простоты изложения): условной базой транзакций для этого эл-та является набор: $\{I_{16}, I_9, I_{14}, I_{10}, I_{13}: 1\}, \{I_{16}, I_9, I_{11}, I_{15}: 1\}$. Эл-ты I_{14}, I_{10}, I_{11}, I_{13} и I_{15} не проходят по мин. уровню поддержки. Условное FP-дерево показано на рисунке 3.

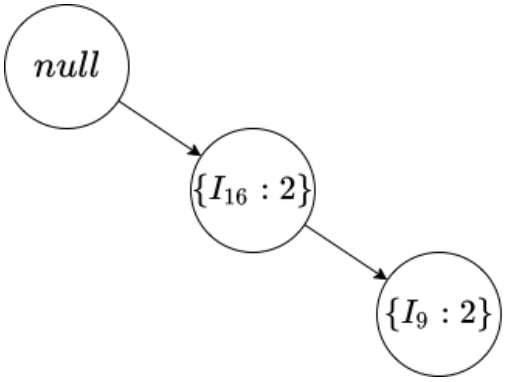

Рисунок 3 – условное FP-дерево для элемента I_4

Соответственно, выбирая по 1 и 2 эл-та из префикса и конкатенируя с I_4, получаем часто встречающиеся подмножества: $\{I_{16}, I_4\}, \{I_9, I_4\}, \{I_{16}, I_9, I_4\}$. По остальным вершинам данные систематизированы в таблице 11.

Таблица 11. Условные FP-деревья и часто встречающиеся подмножества

1-набор	Условная база шаблонов	Условное FP-дерево	Часто встречающиеся подмножества
I_6	$\{I_{11}, I_{12}, I_{14}, I_{15}, I_3: 1\}$ $\{I_{11}, I_{12}, I_{15}: 1\}$	$\langle I_{11}: 2, I_{12}: 2, I_{15}: 2\rangle$	$\{I_{11}, I_6\}, \{I_{12}, I_6\}, \{I_{15}, I_6\}$ $\{I_{11}, I_{12}, I_6\}, \{I_{11}, I_{15}, I_6\}$
I_4	$\{I_{16}, I_9, I_{14}, I_{10}, I_{13}: 1\}$ $\{I_{16}, I_9, I_{11}, I_{15}: 1\}$	$\langle I_{16}: 2, I_9: 2\rangle$	$\{I_{16}, I_4\}, \{I_9, I_4\}, \{I_{16}, I_9, I_4\}$
I_3	$\{I_{16}, I_9, I_{14}, I_1, I_2, I_{10}, I_{13}: 1\}$ $\{I_{11}, I_{12}, I_{14}, I_{15}: 1\}$	$\langle I_{14}: 2\rangle$	$\{I_{14}, I_3\}$
I_{13}	$\{I_{16}, I_9, I_{14}, I_1, I_2, I_{10}: 1\}$ $\{I_{16}, I_9, I_{14}, I_{10}: 1\}$ $\{I_{16}, I_1, I_{10}: 1\}$	$\langle I_{16}: 3, I_9: 2, I_{14}: 2, I_1: 1, I_{10}: 1\rangle$ $\langle I_{16}: 3, I_9: 2, I_{14}: 2, I_{10}: 1\rangle$ $\langle I_{16}: 3, I_1: 1, I_{10}: 1\rangle$	$\{I_1, I_{13}\}, \{I_{14}, I_{13}\}, \{I_9, I_{13}\}$ $\{I_{10}, I_{13}\}, \{I_{16}, I_{13}\}$ $\{I_9, I_{14}, I_{13}\}, \{I_{16}, I_9, I_{13}\}$ $\{I_9, I_{10}, I_{13}\}, \{I_{16}, I_{10}, I_{13}\}$ $\{I_{16}, I_{14}, I_{13}\}, \{I_{14}, I_{10}, I_{13}\}$ $\{I_{16}, I_1, I_{13}\}, \{I_1, I_{10}, I_{13}\}$ $\{I_{16}, I_1, I_{10}, I_{13}\}, \{I_{16}, I_{14}, I_{10}, I_{13}\}$ $\{I_{16}, I_9, I_{10}, I_{13}\}, \{I_{16}, I_{10}, I_{14}, I_{13}\}$ $\{I_{16}, I_9, I_{14}, I_{13}\}$ $\{I_{16}, I_9, I_{14}, I_{10}, I_{13}\}$
I_{10}	$\{I_{16}, I_9, I_{14}, I_1, I_2: 1\}$ $\{I_{16}, I_9, I_{14}: 1\}, \{I_{16}, I_1: 1\}$	$\langle I_{16}: 3, I_9: 2, I_{14}: 2, I_1: 1\rangle$ $\langle I_{16}: 3, I_1: 1\rangle$	$\{I_1, I_{10}\}, \{I_{14}, I_{10}\}, \{I_9, I_{10}\}$ $\{I_{16}, I_{10}\}, \{I_{16}, I_1, I_{10}\}$ $\{I_{16}, I_{14}, I_{10}\}, \{I_9, I_{14}, I_{10}\}$ $\{I_{16}, I_9, I_{10}\}, \{I_{16}, I_9, I_{14}, I_{10}\}$

Окончание таблицы 11

1-набор	Условная база шаблонов	Условное FP-дерево	Часто встречающиеся подмножества
I_8	$\{I_{16}, I_9, I_{11}, I_{12}, I_{14}, I_2 : 1\}$ $\{I_9, I_{11} : 1\}, \{I_{12}, I_{15} : 1\}$	$\langle I_9 : 2, I_{11} : 2, I_{12} : 1\rangle$ $\langle I_{12} : 1\rangle$	$\{I_{12}, I_8\}, \{I_{11}, I_8\}, \{I_9, I_8\}$ $\{I_9, I_{11}, I_8\}$
I_2	$\{I_{16}, I_9, I_{11}, I_{12}, I_{14} : 1\}$ $\{I_{16}, I_9, I_{14}, I_1 : 1\}, \{I_{16}, I_1 : 1\}$	$\langle I_{16} : 3, I_9 : 2, I_{14} : 2, I_1 : 1\rangle$ $\langle I_{16} : 3, I_1 : 1\rangle$	$\{I_9, I_2\}, \{I_{14}, I_2\}, \{I_1, I_2\}$ $\{I_{16}, I_2\}, \{I_{16}, I_{14}, I_2\}$ $\{I_{16}, I_1, I_2\}, \{I_9, I_{14}, I_2\}$ $\{I_{16}, I_9, I_2\}, \{I_{16}, I_9, I_{14}, I_2\}$
I_1	$\{I_{16}, I_9, I_{14} : 1\}, \{I_{16} : 2\}$	$\langle I_{16} : 3\rangle$	$\{I_{16}, I_1\}$
I_{15}	$\{I_{16}, I_9, I_{11} : 1\}, \{I_{11}, I_{12}, I_{14} : 1\}$ $\{I_{11}, I_{12} : 1\}, \{I_{12} : 1\}$	$\langle I_{11} : 3, I_{12} : 2\rangle$ $\langle I_{12} : 1\rangle$	$\{I_{11}, I_{15}\}, \{I_{12}, I_{15}\}$ $\{I_{11}, I_{12}, I_{15}\}$
I_{14}	$\{I_{16}, I_9, I_{11}, I_{12} : 1\}, \{I_{16}, I_9 : 2\}$ $\{I_{11}, I_{12} : 1\}$	$\langle I_{16} : 3, I_9 : 3, I_{11} : 1, I_{12} : 1\rangle$ $\langle I_{11} : 1, I_{12} : 1\rangle$	$\{I_{11}, I_{12}, I_{14}\}, \{I_{11}, I_{14}\},$ $\{I_{12}, I_{14}\}, \{I_9, I_{14}\}, \{I_{16}, I_{14}\}$ $\{I_{16}, I_9, I_{14}\}$
I_{12}	$\{I_{16}, I_9, I_{11} : 1\}, \{I_{11} : 2\}$	$\langle I_{11} : 3\rangle$	$\{I_{11}, I_{12}\}$
I_{11}	$\{I_{16}, I_9 : 2\}, \{I_9 : 1\}$	$\langle I_{16} : 2, I_9 : 2\rangle, \langle I_9 : 1\rangle$	$\{I_{16}, I_{11}\}, \{I_9, I_{11}\}$ $\{I_{16}, I_9, I_{11}\}$
I_9	$\{I_{16} : 4\}$	$\langle I_{16} : 4\rangle$	$\{I_{16}, I_9\}$

Ответ: тут по-хорошему надо выписать все подмножества, которые часто встречаются, но их чрезвычайно много получилось. 1-наборы см. табл. 9, остальные наборы см. табл. 11.

Задание 6

Данные представлены в таблице 12.

Таблица 12. Выборка

x	1	3	5	5	7	7	9	10	8	6
y	9	6	4	5	0	3	1	0	7	8

Построим диаграмму рассеяния точек данных (рис. 4).

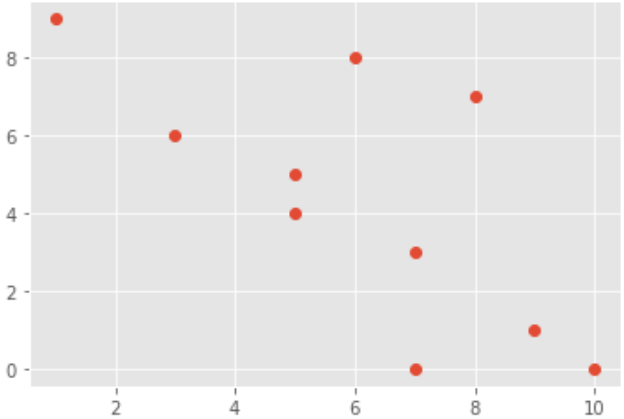

Рисунок 4 – диаграмма рассеяния данных

Выберем в качестве начальных центров кластеров точки (1, 9), (10, 0) и (8, 7). В таблице 13 по шагам расписаны вычисления, связанные с распределением точек по кластерам и обновлением центроидов по алгоритму:

1. Для каждого элемента выборки выделить кластер по формуле $y_i = \underset{y \in Y}{arg\,min}\, \rho\left(x_i, m_y\right)$.
2. Вычисляются новые центроиды:

$$m_y^j = \frac{\sum_{i=1}^{l}[y_i = y]x_i^j}{\sum_{i=1}^{l}[y_i = y]} \tag{8}$$

Таблица 13. Итерации алгоритма k-means

Итерация	Центроиды	Кластеры	Обновленные центроиды
1	(1, 9) (10, 0) (8, 7)	1: (1, 9), (3, 6) 2: (7, 0), (9, 1), (10, 0) 3: (5, 4), (5, 5), (7, 3), (8, 7), (6, 8)	(2, 7.5) (8.67, 0.33) (6.2, 5.4)
2	(2, 7.5) (8.67, 0.33) (6.2, 5.4)	1: (1, 9), (3, 6) 2: (7, 0), (9, 1), (10, 0) 3: (5, 4), (5, 5), (7, 3), (8, 7), (6, 8)	(2, 7.5) (8.67, 0.33) (6.2, 5.4)

На второй итерации центроиды кластеров не поменялись, а значит алгоритм сошелся. Можно визуализировать полученные кластеры (рис. 5).

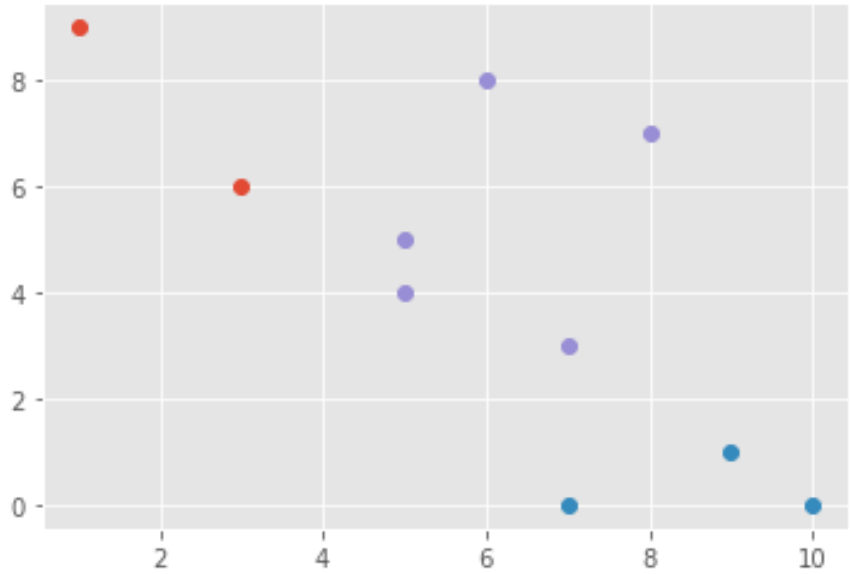

Рисунок 5 – кластеризованные данные

Ответ: кластер 1: (1, 9), (3, 6);
кластер 2: (7, 0), (9, 1), (10, 0);
кластер 3: (5, 4), (5, 5), (7, 3), (8, 7), (6, 8).

Задание 7

Выборка представлена в таблице 14.

Таблица 14. Выборка

x	-9	-6	-5	1	-7	-1	-6	4	6	7
y	6	4	6	1	-4	-5	-8	8	8	4

Построим диаграмму рассеяния (рис. 6).

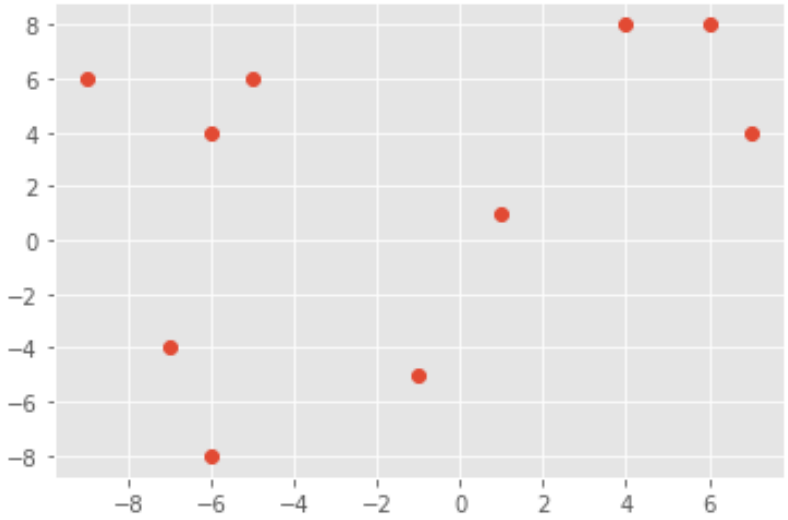

Рисунок 6 – диаграмма рассеяния данных

Алгоритм итеративный, на каждой итерации t выделяются два ближайших кластера и объединяются в один. Расстояние между кластерами вычисляется методом Уорда. Итерации алгоритма показаны в таблице 15.

Таблица 15. Ближайшие кластеры на каждой итерации и расстояния между ними, рассчитанные по методу Уорда

i	Ближайшие кластеры	Расстояние
1	(4, 8) и (6, 8)	$R(W,S) = \frac{1 \cdot 1}{1+1}\rho^2(m_W, m_S) = \frac{1}{2}((4-6)^2 + (8-8)^2) = 2$
2	(-6, 4), (-5, 6)	$R(W,S) = \frac{1 \cdot 1}{1+1}\rho^2(m_W, m_S) = \frac{1}{2}((-6+5)^2 + (4-6)^2) = 2.5$
3	(-7, -4), (-6, -8)	$R(W,S) = \frac{1 \cdot 1}{1+1}\rho^2(m_W, m_S) = \frac{1}{2}((-7+6)^2 + (-4+6)^2) = 8.5$
4	(-9, 6), [(-6, 4), (-5, 6)]	$R(W,S) = \frac{1 \cdot 2}{1+2}\rho^2(m_W, m_S) = \frac{2}{3}((-9+5.5)^2 + (6-5)^2) = 8.8(3)$

Окончание таблицы 15

i	Ближайшие кластеры	Расстояние
5	(7, 4), [(4, 8), (6, 8)]	13. (3)[1]
6	(1, 1), (-1, -5)	20
7	[(-7, -4), (-6, -8)] [(1, 1), (-1, -5)]	58.25
8	[(-9, 6), (-6, 4), (-5, 6)] [(-7, -4), (-6, -8), (1, 1), (-1, -5)]	169.3452…
9	[(7, 4), (4, 8), (6, 8)] [(-9, 6), (-6, 4), (-5, 6), (-7, -4), (-6, -8), (1, 1), (-1, -5)]	319.6381…

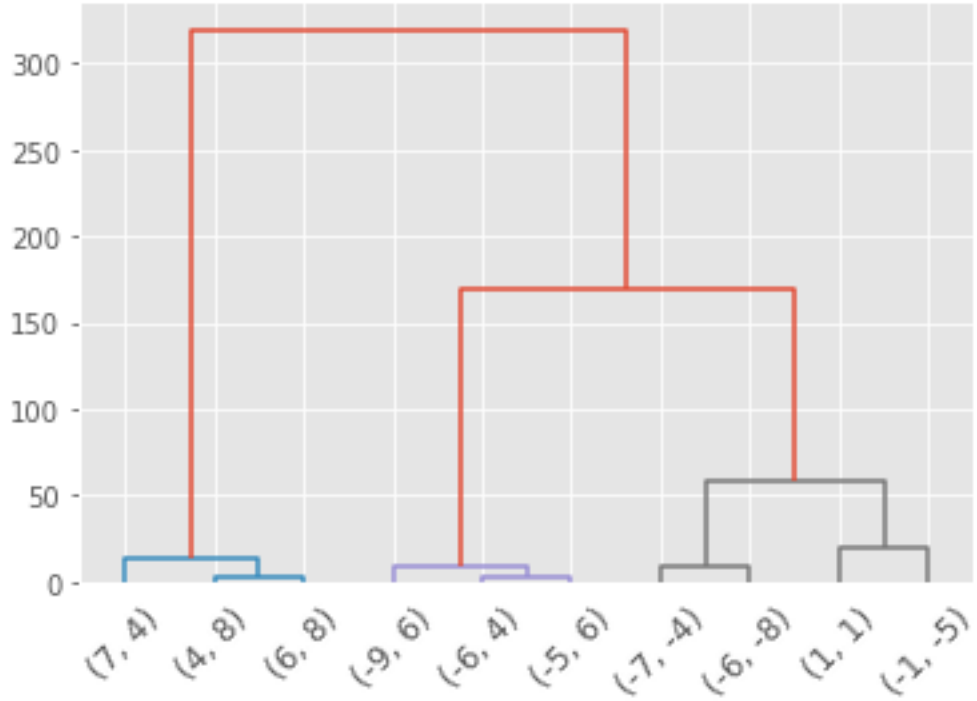

Рисунок 7 - дендрограмма

На дендрограмме явственно видно, что происходит резкий скачок расстояний объединяемых кластеров после объединения в 3 кластера, то есть оптимальное количество кластеров равно 3. Данные кластеры показаны разными цветами.

Стоит отметить, что если базироваться на критерии максимума разницы расстояний сливаемых на соседних итерациях кластеров, то выделятся 2 кластера. В то же время, можно построить т.н. «каменистую осыпь», на которой резкий подъем расстояний будет заметен после выделения 5 кластеров. Однако все методы оценки количества кластеров являются нестрогими и по большей части «субъективными», что позволяет каждый конкретный случай трактовать наиболее

[1] Здесь и далее в этом столбце вычисления абсолютно аналогичны и чисто механические.

подходящим образом. В данном случае визуально по дендрограмме, как сказано выше, можно выделить три кластера.

Ответ: дендрограмму см. на рис. 7.

Кластер 1: (7, 4), (4, 8), (6, 8)

Кластер 2: (-9, 6), (-6, 4), (-5, 6)

Кластер 3: (-7, -4), (-6, -8), (1, 1), (-1, -5)

Задание 8

Длина выборки равна 10, $\pi = 0.25$, а значит в 10-окрестности точки должно лежать по меньшей мере 3 точки ($10 \cdot 0.25 = 2.5$, вот больше или столько же точек должно быть), чтобы она не была (r, π)-аномалией.

В соответствии с алгоритмом вложенных циклов для каждой точки напрямую подсчитываются количества других точек из ее r-окрестности. В алгоритме предусмотрен выход из цикла при достижении $\pi \cdot n$ элементов, однако в таблице 16 приведено *полное* количество точек из ее r-окрестности для консистентности.

Пример расчета для точки (-9, 6):

- (-6, 4): расстояние ~ 3.6 (count = 1)
- (-5, 6): расстояние 4 (count = 2)
- (1, 1): расстояние ~ 11.2
- (-7, -4): расстояние ~ 10.2
- (-1, -5): расстояние ~ 13.6
- (-6, -8): расстояние ~ 14.3
- (4, 8): расстояние ~ 13.2
- (6, 8): расстояние ~ 15.1
- (7, 4): расстояние ~ 16.1

Таким образом, (-9, 6) – (10, 0.25)-аномалия.

Теперь расчет для точки (-6, 4):

- (-9, 6) расстояние ~ 3.6 (count = 1)
- (-5, 6) расстояние ~ 2.2 (count = 2)
- (1, 1) расстояние ~ 7.6 (count = 3)

На этом этапе обработка точки закончена согласно алгоритму, так как count ≥ 2.5.

Таблица 16. Количество точек в 10-окрестности элементов выборки

x	-9	-6	-5	1	-7	-1	-6	4	6	7
y	6	4	6	1	-4	-5	-8	8	8	4
n	2	4	4	7	4	3	2	4	3	3

Ответ: (-9, 6), (-6, -8)