Тема 4. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

4.1. Введение в корреляционный анализ

Корреляционный анализ применяют для оценки зависимости выходных полей данных от входных факторов и устранения незначащих факторов. Принцип корреляционного анализа состоит в поиске таких значений, которые в наименьшей степени коррелированы (взаимосвязаны) с выходным результатом. Такие факторы могут быть исключены из результирующего набора данных практически без потери полезной информации. Критерием принятия решения об исключении является порог значимости. Если модуль корреляции (степень взаимозависимости) между входным и выходным факторами меньше порога значимости, то соответствующий фактор отбрасывается как незначащий.

Модуль коэффициента свидетельствует о степени зависимости: чем ближе его значение к 0, тем слабее линейная зависимость. Чем ближе коэффициент корреляции от 0 к 1, тем сильнее прямая линейная зависимость, чем ближе от 0 к — 1, тем сильнее обратная линейная зависимость. На практике считается, что если модуль коэффициента корреляции больше 0,6, то линейная зависимость сильная, а если менее 0,3, то почти отсутствует.

Однако, низкая степень корреляции между входным и прогнозируемым полями не означает отсутствие других, нелинейных зависимостей. Кроме того, при построении линейных моделей стоит рассмотреть такой входной фактор внимательнее, так как он может быть использован для проектирования признаков.

Методами корреляционного анализа решаются следующие задачи:

- *взаимосвязь*. Установление наличия зависимости между двумя признаками и определение её силы;
- *прогнозирование*. Предсказание поведения одного признака на основе изменения другого, коррелирующего с первым;
- *отбор переменных*. Корреляционный анализ позволяет производить выбор набора входных переменных для аналитической модели в наименьшей степени коррелирующих между собой и в наибольшей степени коррелирующих с выходной переменной. Это позволяет сделать работу аналитических моделей более точной и устойчивой.

Для расчета коэффициентов корреляции используются четыре метода:

- *коэффициент корреляции Пирсона* — с его помощью можно определить силу и направление линейной зависимости между двумя процессами, происходящими одновременно;
- *коэффициент Tau-b Кендалла* — коэффициент ранговой корреляции, применяется для выявления количественной взаимосвязи между переменными, если их можно ранжировать. Рекомендуется использовать для категориальных данных;

- *экстремум взаимнокорреляционной функции* — вычисляет максимальный по модулю из коэффициентов корреляции двух процессов, рассчитанных при всевозможных временных сдвигах. Следует применять, если необходимо узнать линейную зависимость между двумя процессами или частями процессов, происходящими с некоторым временным лагом;
- *коэффициент корреляции Спирмена* — еще один вариант ранговой корреляции. Для числовых полей для оценки силы связи используются не численные значения, а соответствующие им ранги. Поэтому для любых монотонных последовательностей коэффициент Спирмена будет равен –1 или 1.

4.2. Методические указания

В файле *Задача 4.1. Потребительские цены.xlsx* имеются данные о ежемесячной потребительской цене (тыс. руб.) на четыре товара за 12 мес. (рис. 4.1).

	A	B	C	D	E
1	Месяц	Товар 1	Товар 2	Товар 3	Товар 4
2	Январь	10	20	15	25
12	Ноябрь	17	25	7	25
13	Декабрь	13	23	10	24

Рис. 4.1

Требуется определить товары-заменители и сопутствующие товары, имея временные ряды объемов продаж. У товаров-заменителей должна быть большая отрицательная корреляция, так как увеличение продаж одного товара ведет к спаду продаж второго, а у сопутствующих товаров — большая положительная корреляция.

Решение

Создадим новый пакет *Корреляционный анализ*. Выполним импорт исходных данных. Для этого создадим узел сценария, выполняющий действие импорта (рис. 4.2).

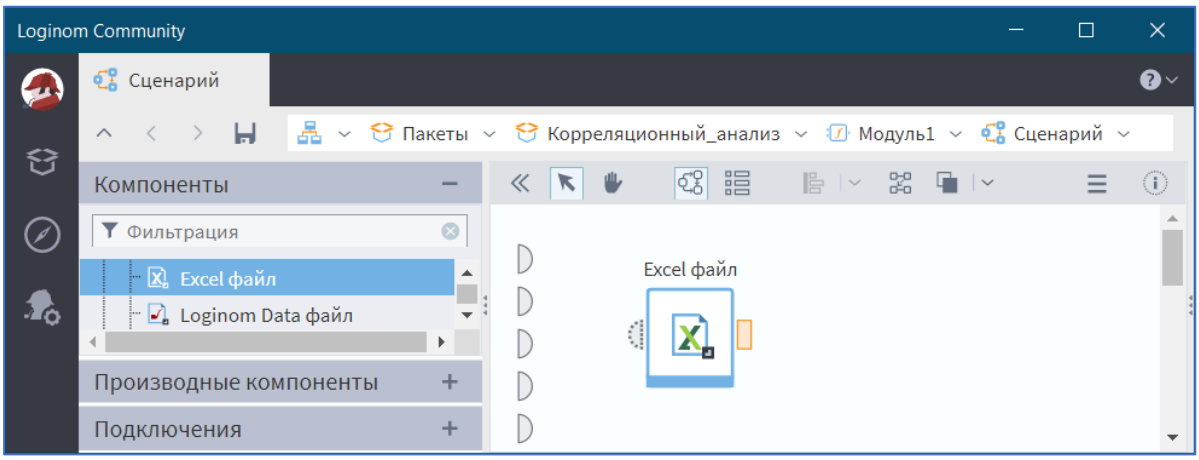


Рис. 4.2

Вызовем *Мастер настройки.* Пройдем шаги мастера, указав в описании узла метку *Задача 4.1. Потребительские цены.*

Добавим визуализатор *Таблица* к узлу сценария (рис. 4.3).

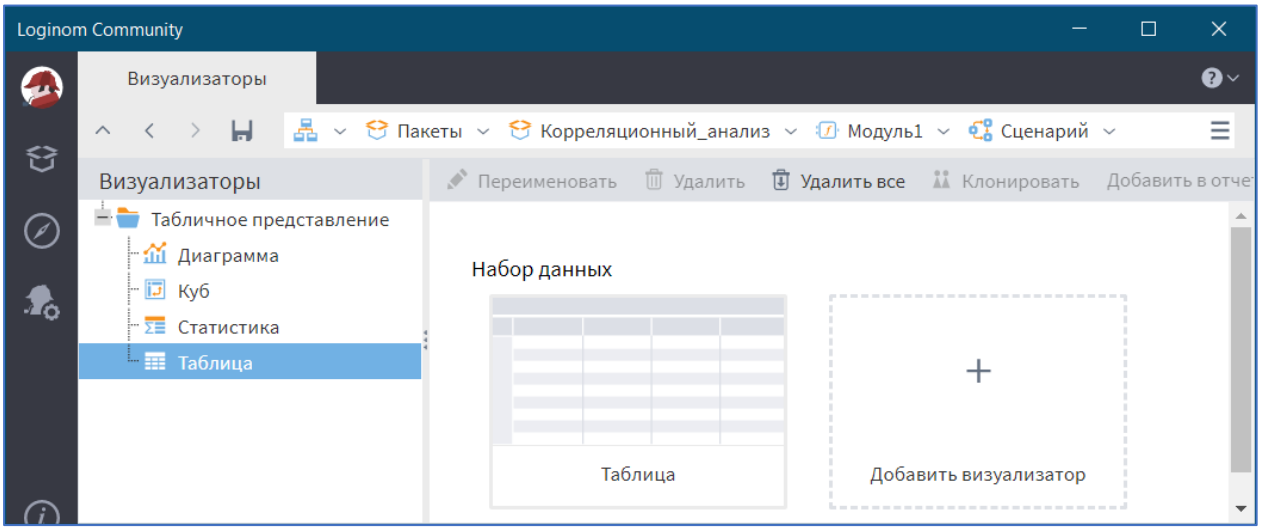
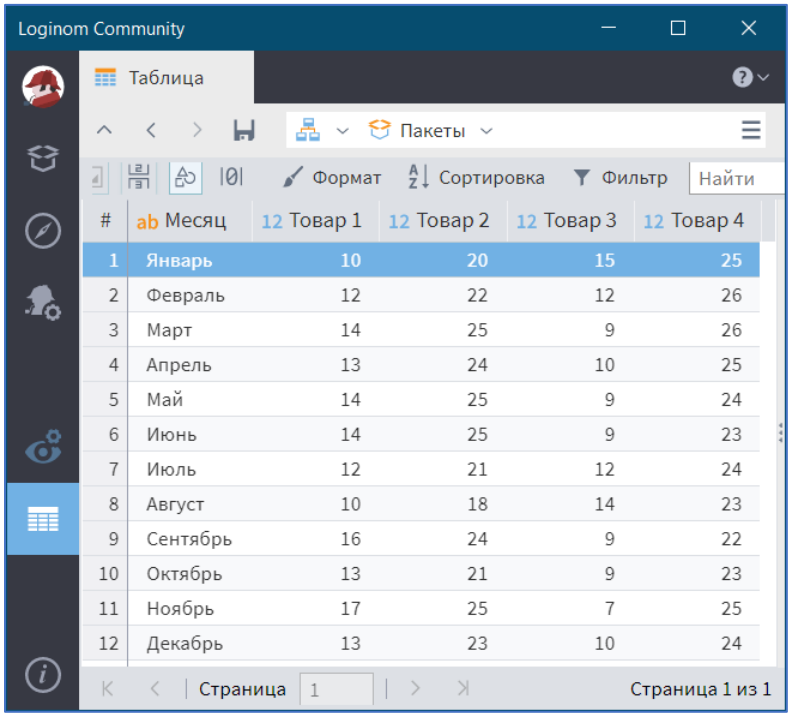


Рис. 4.3

Таблица с исходными данными имеет вид (рис. 4.4).



#	Месяц	Товар 1	Товар 2	Товар 3	Товар 4
1	Январь	10	20	15	25
2	Февраль	12	22	12	26
3	Март	14	25	9	26
4	Апрель	13	24	10	25
5	Май	14	25	9	24
6	Июнь	14	25	9	23
7	Июль	12	21	12	24
8	Август	10	18	14	23
9	Сентябрь	16	24	9	22
10	Октябрь	13	21	9	23
11	Ноябрь	17	25	7	25
12	Декабрь	13	23	10	24

Рис. 4.4

Для выявления товаров-заменителей проведем корреляционный анализ на основе импортированных данных. Для этого переместим компонент *Корреляционный анализ* в рабочую область сценария. Последовательность обработки данных задается соединением выходного порта узла импорта с входным портом корреляционного анализа (рис. 4.5).

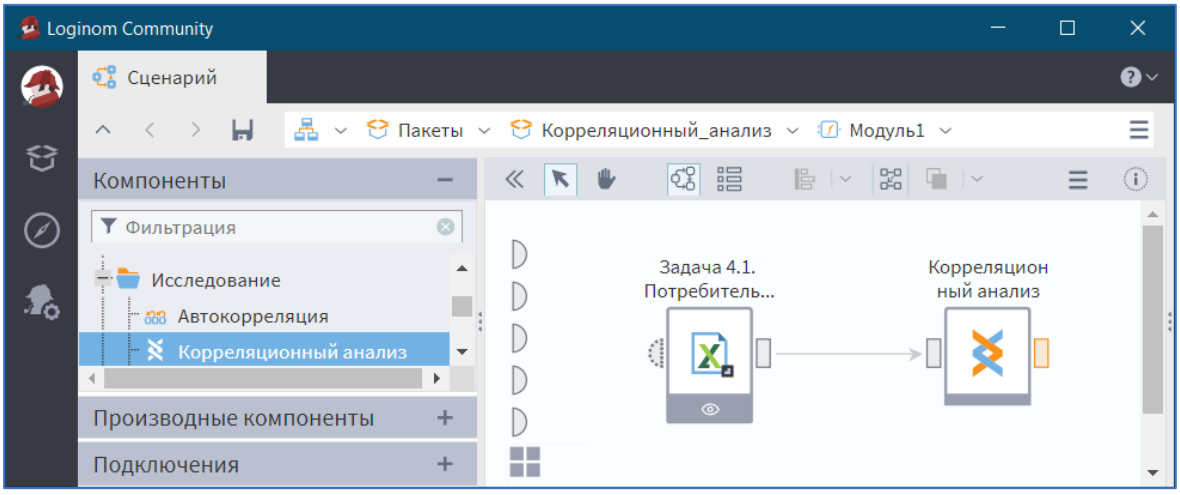


Рис. 4.5

Пройдем шаги *Мастера настройки*. На шаге 1 настроим параметры столбцов. Поскольку надо определить степень зависимости между продажами первого товара и остальных товаров, первый товар определим как *Набор 1*, а остальные товары – как *Набор 2* (рис. 4.6).

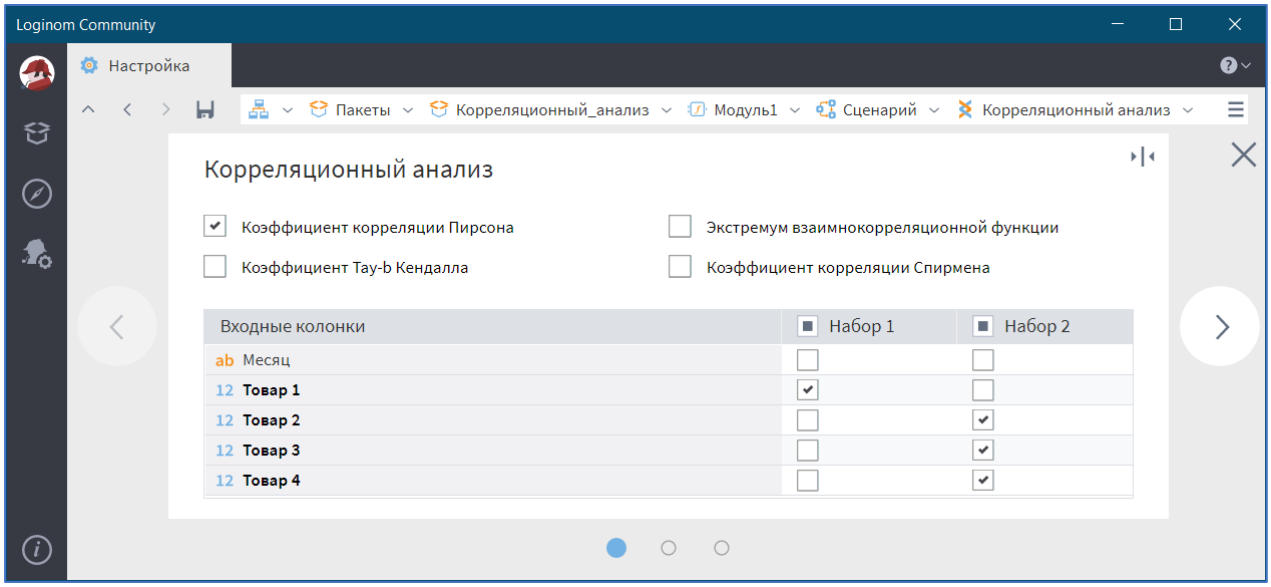


Рис. 4.6

Добавим визуализатор *Таблица* к узлу сценария (рис. 4.7).

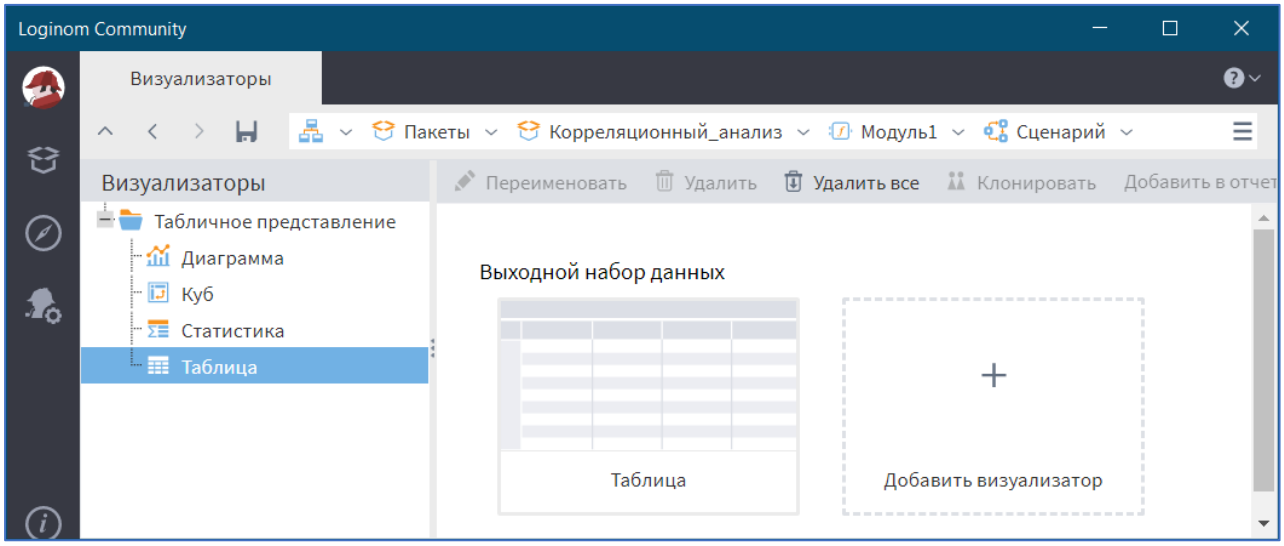


Рис. 4.7

В визуализаторе будут представлены результаты корреляционного анализа (рис. 4.8).

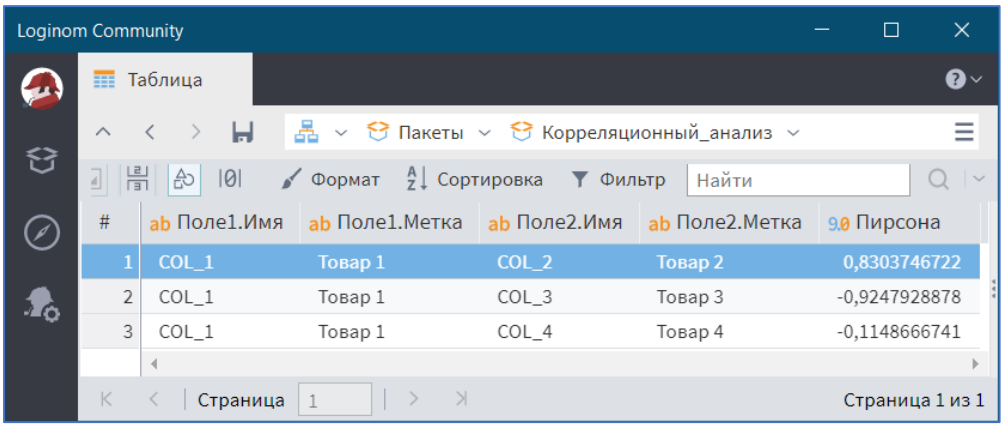


Рис. 4.8

Как видно из таблицы, ряд продаж для второго товара имеет очень большую положительную, а третьего товара — отрицательную корреляцию. Из этого можно сделать вывод, что второй товар, возможно, является сопутствующим товаром, а третий товар — заменителем первого товара. Корреляция с продажами четвертого товара с первым является отрицательной, но при этом абсолютное значение корреляции невелико, и, следовательно, можно говорить об отсутствии взаимосвязи между продажами первого и четвертого товаров.

4.3. Задание для самостоятельной работы

В файле *Задача 4.2. Производительность труда.xlsx* приведены данные обследования 17 предприятий с целью выявления факторов, влияющих на производительность труда (рис. 4.9). Переменные y_1 и y_2 являются результативными, а x_1–x_{25} — факторными.

	A	B	C	D	E	AA	AB
1	Номер	y1 – выработка на одного работника, тыс. руб.	y2 – выработка на одного рабочего, тыс. руб.	x1 – доля рабочих, занятых наблюдением за работой автоматов, %	x2 – доля рабочих, занятых при машинах и механизмах, %	x24 – доля автоматов в технологическом оборудовании	x25 – доля в технологическом оборудовании автоматических линий
2	1	71,05	86,31	10,1	35,6	33,2	17,9
3	2	62,68	80,95	2,6	41,3	32,8	2,5
17	16	39,66	50,81	6,9	37,1	44,2	0,2
18	17	50,52	65,51	17,9	21,6	73,2	3

Рис. 4.9

Требуется определить факторы, влияющие на объем выработки на одного работника и на одного рабочего.