Представление категорийных данных в виде таблиц и диаграмм

В предыдущей заметке таблицы и диаграммы применялись для представления числовых данных. Однако часто данные носят не числовой, а категориальный характер. В этой заметке изучаются способы организации и представления категорийных данных в виде таблиц и диаграмм.[1]

Вернемся к анализу доходности взаимных фондов. Кроме среднегодовой доходности фонды характеризуются риском, связанном с инвестированием в эти фонды. Взаимные фонды могут иметь очень низкий, низкий, средний, высокий и очень высокий риск. При работе с категорийными переменными данные сначала заносятся в сводную таблицу, а затем графически представляются в виде гистограмм, круговых диаграмм или диаграмм Парето.

Сводная таблица

По внешнему виду сводная таблица для категорийных данных напоминает распределение частот для числовых данных. Чтобы проиллюстрировать процесс ее построения, рассмотрим данные о классификации взаимных фондов по уровню риска (рис. 1).

Уровень риска	Количество фондов	Процентная доля фондов
Очень низкий	6	2,3%
Низкий	76	29,3%
Средний	82	31,7%
Высокий	80	30,9%
Очень высокий	15	5,8%
	259	100,0%

Рис. 1. Уровень риска 259 взаимных фондов. Частоты и процентные доли

Линейчатая диаграмма

Информацию, содержащуюся в таблице (рис. 1), можно представить в виде линейчатой диаграммы (рис. 2), в которой каждая категория элементов изображается в виде столбца. Высота столбца равна частоте или процентной доле элементов выборки, относящихся к данной категории.

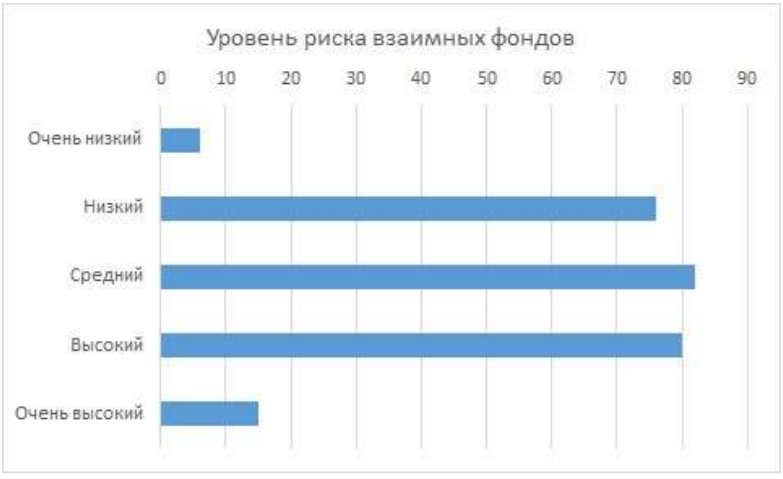

Рис. 2. Линейчатая диаграмма, отображающая уровень риска фондов

Круговая диаграмма

Существует еще один весьма популярный способ отображения информации, содержащейся в сводной таблице, — круговая диаграмма (рис. 3). При построении круговых диаграмм используется тот факт, что угол окружности равен 360°. Круг разделяется на секторы, углы которых соответствуют процентным долям каждой категории. Например, на рис. 3 показан сектор, соответствующий доле взаимных фондов с низким риском, которая равна 29,3%. При построении круговой диаграммы величина 360° умножается на 0,293. В результате образуется сектор, угол которого равен 105,6°. Как видим, круговая диаграмма позволяет отразить долю каждой категории в общем «пироге».

[1] Используются материалы книги Левин и др. Статистика для менеджеров. – М.: Вильямс, 2004. – с. 124–138

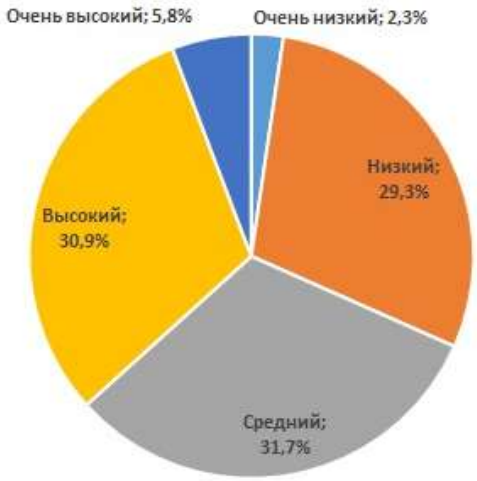

Рис. 3. Круговая диаграмма, отображающая уровень риска фондов

Цель графического представления данных — точность и ясность. Например, рис. 2 и 3 отображают одинаковую информацию. Какой из двух видов диаграмм предпочесть — дело вкуса. В частности, некоторые исследования показывают, что люди труднее воспринимают круговые диаграммы. Оказывается, человеку намного проще интерпретировать разницу между высотами столбцов в линейчатых диаграммах, чем углы секторов в круговых диаграммах. Обратите внимание на то, что по рис. 3 нелегко определить, какая из категорий фондов больше — с низким, средним или высоким уровнем риска. В то же время по линейчатой диаграмме легко определить, что доля фондов со средним уровнем риска больше, чем доли фондов с высоким и низким уровнями риска.

С другой стороны, круговые диаграммы четко демонстрируют, что сумма долей всех категорий равна 100%. Таким образом, выбор диаграммы является субъективным и часто зависит от предпочтений пользователя. Если необходимо сравнивать доли категорий, лучше применять линейчатые диаграммы. Если важно продемонстрировать вклад долей отдельных категорий в общий «пирог», лучше использовать круговые диаграммы.

Диаграмма Парето

Существует более информативный способ графического изображения категорийных данных — диаграмма Парето. Она особенно полезна, если количество категорий велико. Диаграмма Парето — это особая разновидность вертикальной диаграммы, в которой категории приводятся в порядке убывания их частот одновременно с полигоном накопленных частот. Это позволяет выделить наиболее важные категории из большого количества менее значимых групп. Диаграмма Парето получила широкое распространение при анализе производственных процессов и контроле качества (см., например, АВС-анализ и принцип Парето для бизнеса).

Например, для построения Диаграммы Парето на основе данных рис. 1, необходимо отсортировать строки по убыванию, и одновременно отобразить как количество фондов в каждой категории, так и интегральный процент (рис. 4).

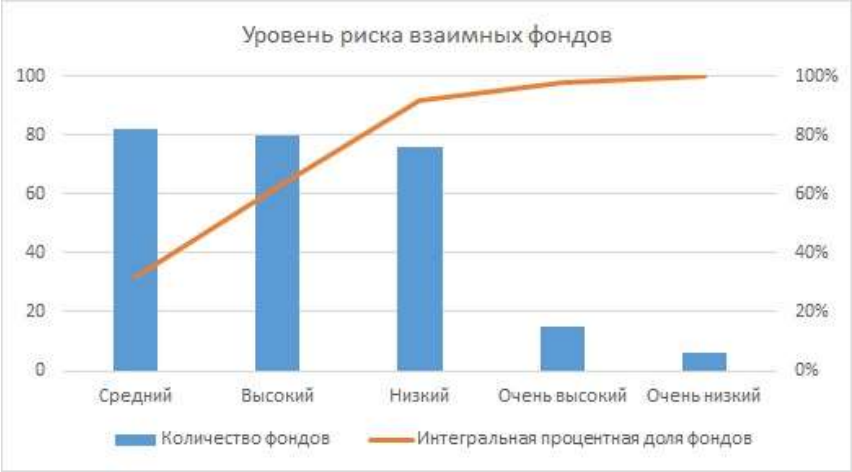

Рис. 4. Диаграмма Парето, отображающая специфику фондов

Надо отметить, что в Excel2013 предоставляется стандартная возможность построения таких комбинированных диаграмм (рис. 5). Если же у вас Excel2007, то вам придется помучиться (см., например, Диаграмма Excel с двумя осями ординат).

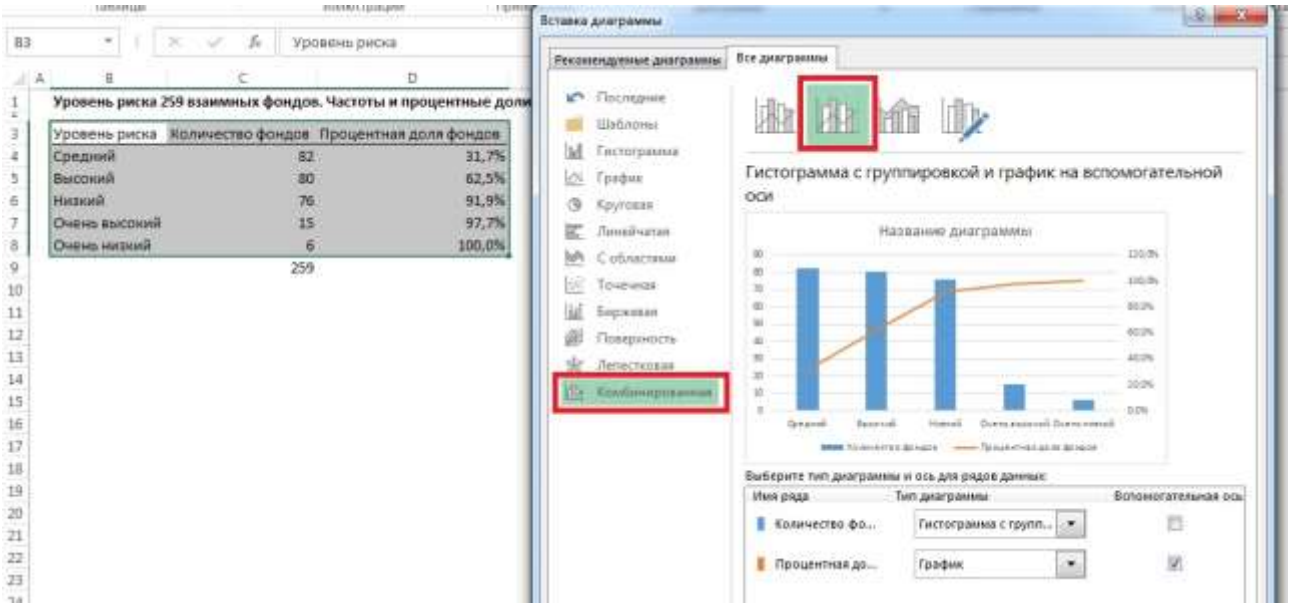

Рис. 5. Построение комбинированной диаграммы

Представление двумерных категорийных данных

Довольно часто необходимо анализировать пары категорийных переменных. Для этого используют таблицы сопряженности признаков и нормированные диаграммы.

Таблица сопряженности признаков. Чтобы можно было одновременно анализировать две категорийные переменные, образующие пару, используются таблицы перекрестной классификации с двумя входами, или таблицы сопряженности признаков (их также называют факторными таблицами). Например, может возникнуть вопрос: существует ли зависимость между уровнем риска и платой, взимаемой фондами за осуществление продаж своих акций (рис. 6)?

	Уровень риска					
Плата	**Очень высокий**	**Высокий**	**Средний**	**Низкий**	**Очень низкий**	**Всего**
Да	4	35	23	31	2	95
Нет	11	45	59	45	4	164
Всего	15	80	82	76	6	259

Рис. 6. Таблица сопряженности признаков, содержащая данные об уровне риска и плате, взимаемой фондами за осуществление продаж своих акций

Чтобы выявить возможную зависимость между специализацией фонда и прейскурантом его комиссионных сборов, эти результаты сначала преобразуют в процентные доли, используя следующие три базиса:

a) общую сумму (259 взаимных фондов);
b) сумму по строкам (фонды, взимающие плату за продажу своих акций, и фонды без брокерской комиссии);
c) сумму по столбцам (пять уровней риска).

Удобную возможность построения таблиц сопряжения дает опция Excel *Сводные таблицы*. Для начала нужно представить исходные данные в виде строк, в каждой из которых содержатся все исследуемые параметры (рис. 7). Далее выделяем область B3:D13, и проходим по меню Вставка → Сводная таблица. В открывшемся окне *Создание сводной таблицы* указываем *на существующий лист* и в поле *Диапазон* кликаем на ячейку, где мы хотели расположить левый верхний угол сводной таблицы, кликаем Ok. (Если вы хотите разместить сводную таблицу на отдельном листе, сразу после открытия окна *Создание сводной таблицы*, кликните Ok.)

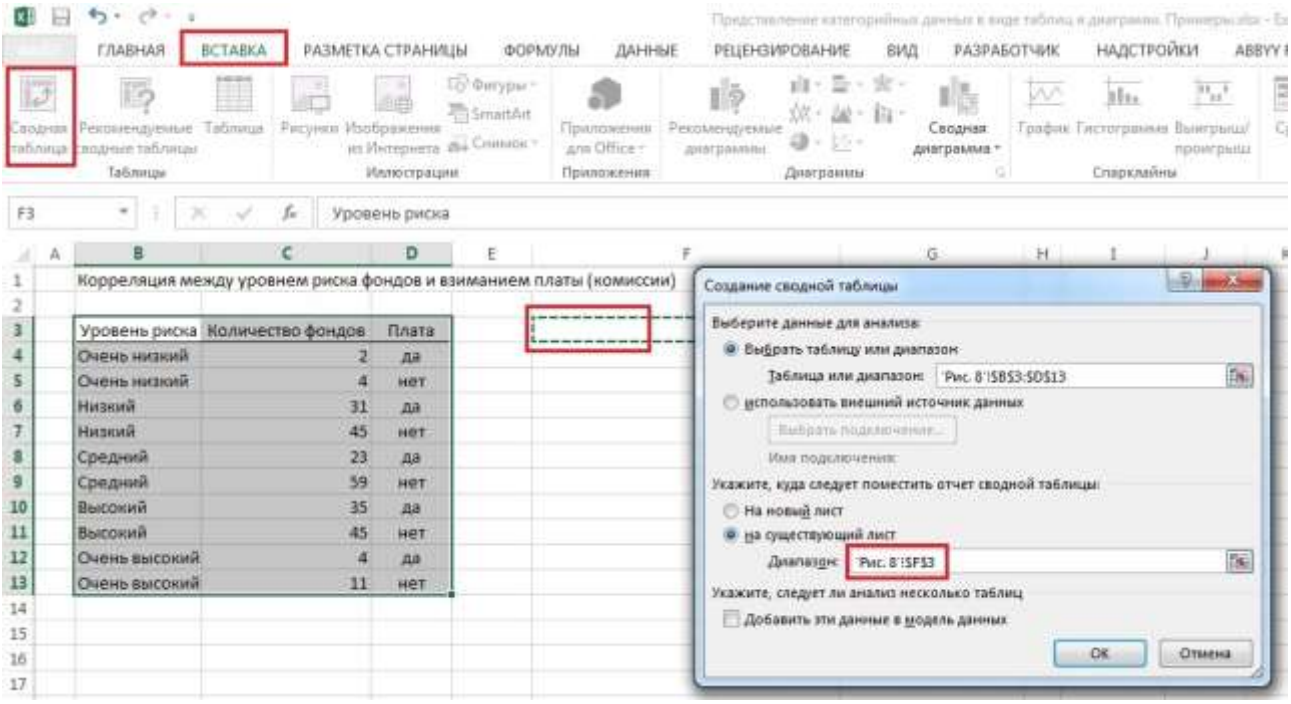

Рис. 7. Построение сводной таблицы

Для настройки сводной таблицы просто перетащите строки из верхней части области *Поля сводной таблицы* в нижнюю, как указано на рис. 8.

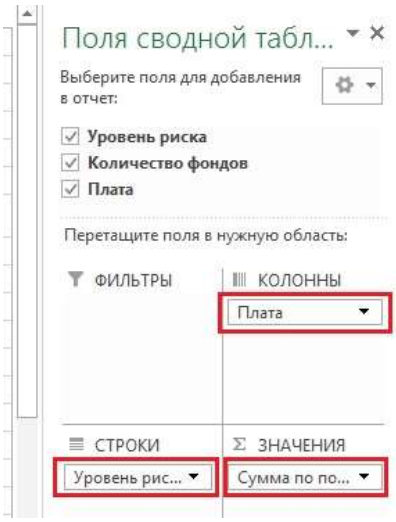

Рис. 8. Настройка полей сводной таблицы

Логично расположить строки в сводной таблицы в порядке возрастания (или убывания) степени риска. Для этого надо по очереди выбрать каждую строку, выбрав ячейку в области *Название строк* (например, *Очень высокий*), кликнуть правой кнопкой мыши, и выбрать в контекстном меню Переместить, указав, куда именно переместить выбранную строку (рис. 9).

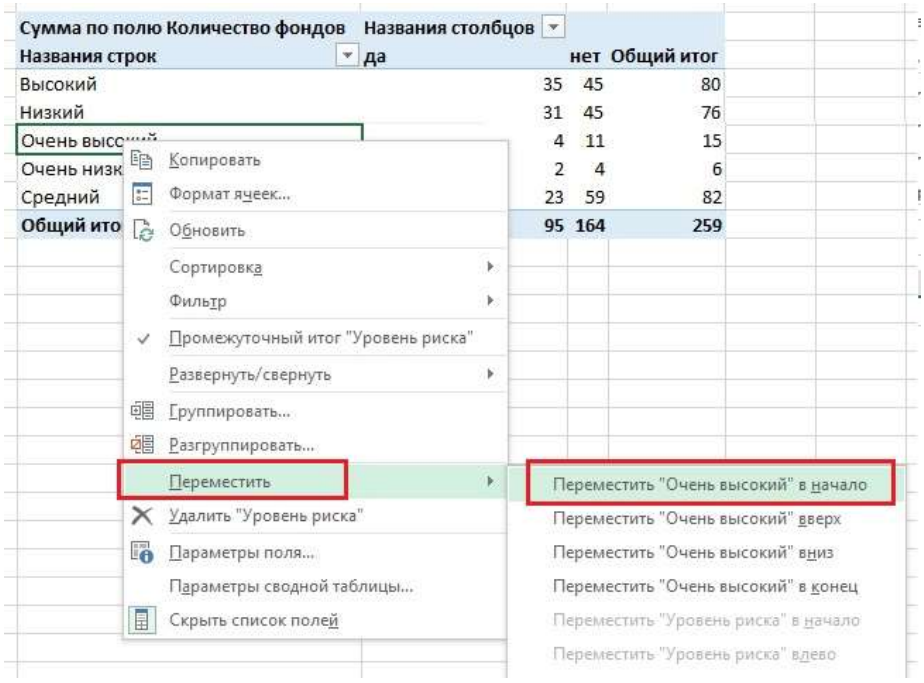

Рис. 9. Перетаскивание строк сводной таблицы

И, наконец, мы можем выбрать базис для анализа процентных долей. Встаньет в любую ячейку в области значений (рис. 10), кликните правой кнопкой мыши, и в открывшемся контекстном меню выберите *Параметры полей значений*.

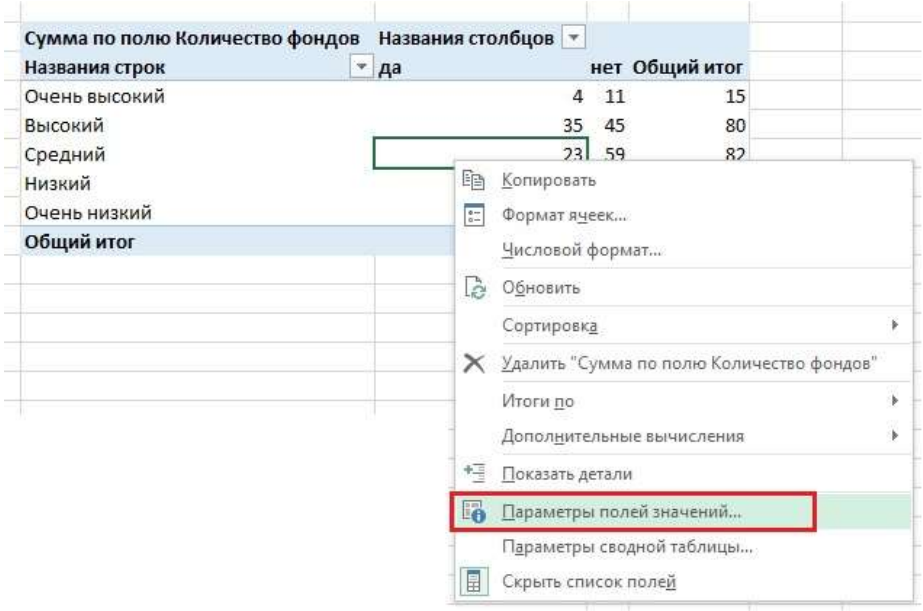

Рис. 10. Параметр поля значений

В окне *Параметры полей значений* перейдите на закладку Дополнительные вычисления, и выберите одну из опций (рис. 11):

- % от общей суммы
- % от суммы по столбцу
- % от суммы по строке

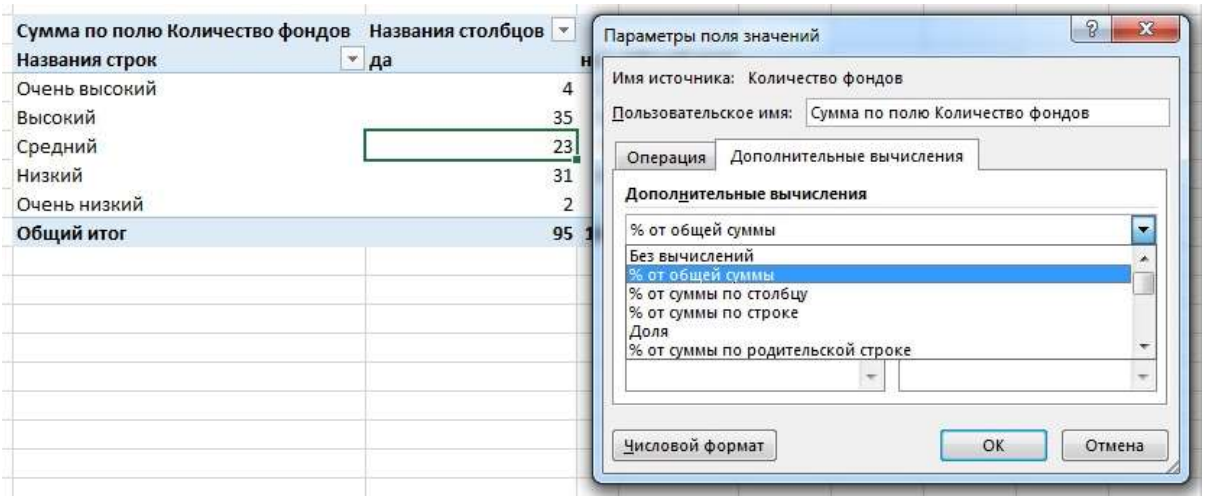

Рис. 11. Выбор базиса процентной доли

Поскольку нас интересует корреляция между степенью риска и наличием комиссии, уместно выбрать опцию % от суммы по строке. Мы увидим, подчиняется ли закономерности доля фондов, взимающих комиссию, при переходе от фондов с очень высоким риском к фондам с очень низким риском (рис. 12). Явной тенденции обнаружить не удалось.[2]

Сумма по полю Количество фондов	Названия столбцов	
Названия строк	да	нет
Очень высокий	27%	73%
Высокий	44%	56%
Средний	28%	72%
Низкий	41%	59%
Очень низкий	33%	67%

Рис. 12. Доля фондов, взимающих комиссию по уровням риска

Нормированные диаграммы

Для визуализации двумерных категорийных данных часто строят нормированные диаграммы, то есть диаграммы, в которых высота столбиков равна 1 (100%) вне зависимости от общего числа случаев в той или иной категории. На рис. 13 представлен пример такой диаграммы. Четко видна закономерность: доля трафика google выросла летом – осенью 2012 г. с 40 до 55%, а затем вновь упала до 40% (для меня остается загадкой, с чем это связано ☺).

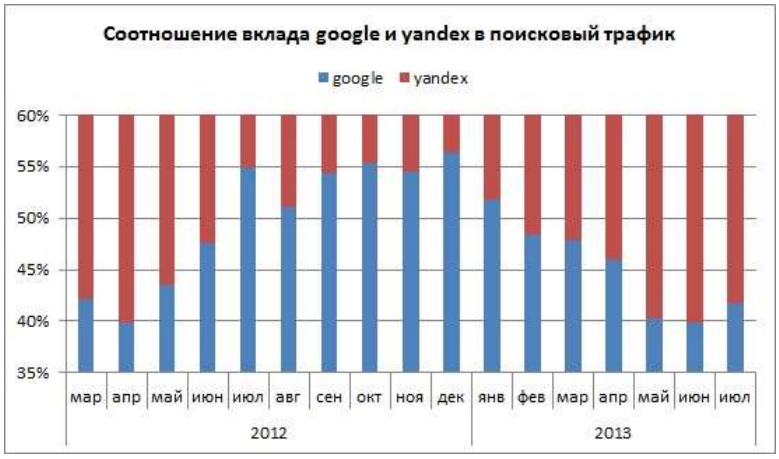

Рис. 13. Соотношение вклада google и yandex в поисковый трафик сайта baguzin.ru

Предыдущая заметка Представление категорийных данных в виде таблиц и диаграмм

Следующая заметка Искусство графического представления данных

К оглавлению Статистика для менеджеров с использованием Microsoft Excel

[2] Любопытно, что авторы книги такую закономерность (на тех же исходных данных) увидели ☺