Санкт-Петербургский
государственный
университет
www.spbu.ru

Анализ данных

Виды данных

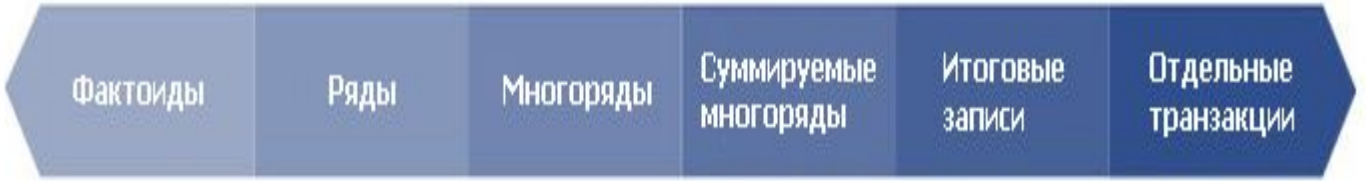
Графеева Н.Г.

2017

Анализ данных. Виды данных

Основные виды данных

Данные, представляемые для анализа, могут быть весьма разнообразного вида: от простых фактоидов (результатов чьего-то анализа) до «сырых» транзакций, изучение которых целиком и полностью является задачей аналитика.



Анализ данных. Виды данных

Фактоид

Фактоид – это часть общей информации. **Фактоид** рассчитывается из исходных (сырых) данных и акцентирует внимание на конкретной детали.

Пример: 36.7% кофе в 2000 году потребили женщины.

Анализ данных. Виды данных

Ряд (series)

Ряд - это когда один вид информации (зависимая переменная) сопоставляется другому виду информации (независимая переменная). Информация, соответствующая зависимой переменной может носить агрегированный характер.

Температура воды °C(°F)	Время получения ожога 1 степени
46.7 (116)	35 минут
50 (122)	1 минута
55 (131)	5 секунд
60 (140)	2 секунды
65 (149)	1 секунда
67.8 (154)	мгновенно

В примере независимая переменная – температура воды, зависимая переменная – время, необходимое взрослому человеку для получения ожога 1 степени

Анализ данных. Виды данных

Временной ряд (time series)

Ряд называется **временным**, если в качестве независимой переменной выступает время.

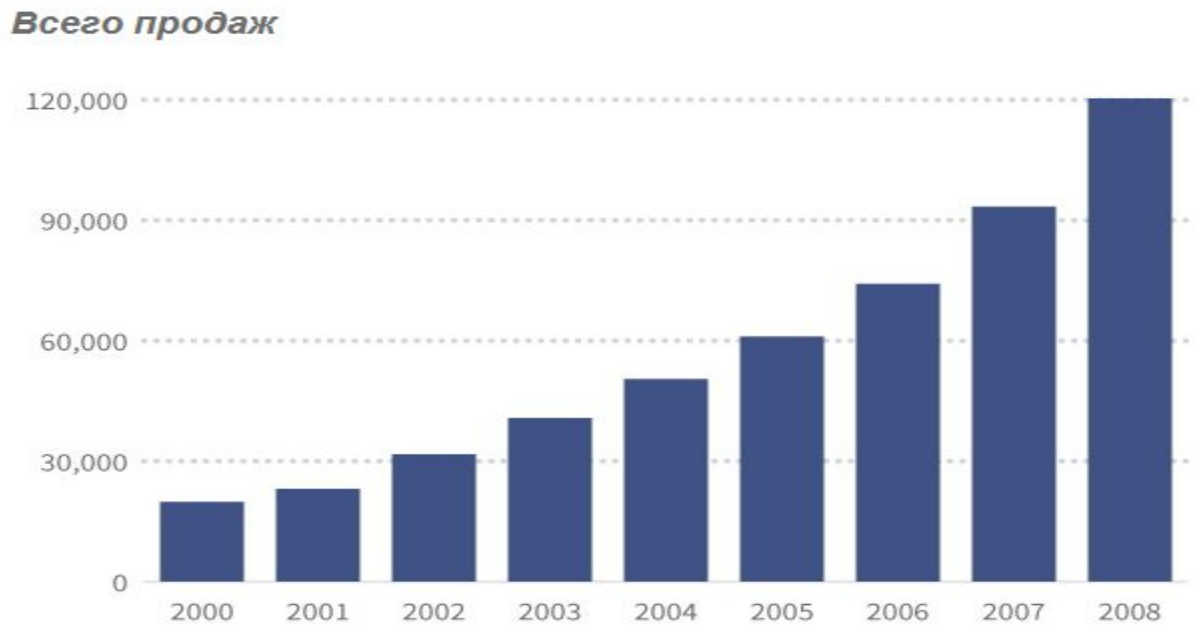
Год	2000	2001	2002	2003
Всего продано	19795	23005	31711	40728

В этом примере общее количество проданного кофе зависит от года. Поэтому год − это независимая переменная («выберите год, любой год»), а количество продаж − зависимая («в этом году потребление кофе составляет 23,005 чашек»).

Анализ данных. Виды данных

Визуализация рядов

Ряды удобно отображать в виде столбчатой диаграммы:



Анализ данных. Виды данных

Многоряды

В **многорядных** данных есть несколько единиц зависимой информации и одна единица независимой информации. Расширенный пример с ожогами:

Температура воды °C, (°F)	Время получения ожога 1 степени	Время получения ожога 2 и 3 степени
46.7 (116)	35 минут	45 минут
50 (122)	1 минута	5 минут
55 (131)	5 секунд	25 секунд
60 (140)	2 секунды	5 секунд
65 (149)	1 секунда	2 секунды
67.8 (154)	мгновенно	1 секунда

Здесь температура – независимая переменная, ожоги (1, 2 и 3 степени) – зависимая.

Анализ данных. Виды данных

Многоряды (пример с кофе)

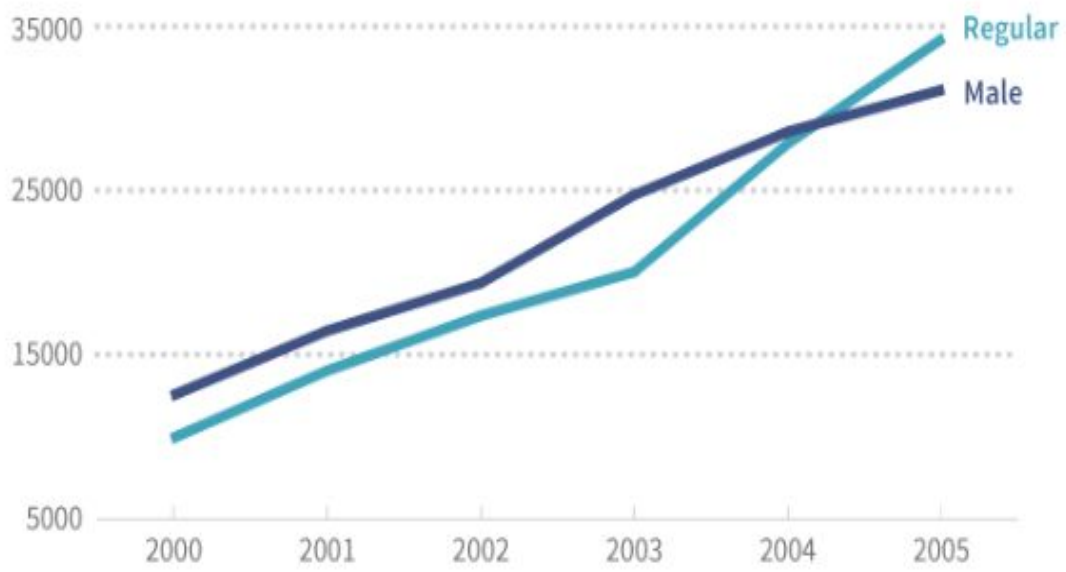
Год	2000	2001	2002	2003	2004	2005
Мужчины	12534	16452	19362	24726	28567	31110
Обыкновенный	9929	14021	17364	20035	27854	34201

С таким набором данных мы знаем несколько фактов, например, о 2001 годе. Мы знаем, что 16452 чашек было продано мужчинам, и что было продано 14021 чашка обычного кофе (с кофеином, сливками/молоком и сахаром). Однако мы не знаем, как объединить эти данные в практических целях: они абсолютно не связаны между собой. Мы не можем сказать, какой процент обычного кофе был продан мужчинам или сколько чашек досталось женщинам.

Анализ данных. Виды данных

Визуализация многорядов

Мы можем показывать **многоряды** вместе, но не можем проагрегировать или объединить их так, чтобы это имело смысл.



Анализ данных. Виды данных

Суммируемые многоряды

Как следует из названия, **суммируемые многоряды** – это отдельный показатель (пол, вид кофе), разбитый на подгруппы.

Год	2000	2001	2002	2003	2004	2005	2006	2007	2008
Мужчины	12534	16452	19362	24726	28567	31110	39001	48710	61291
Женщины	7261	6553	12349	16002	21873	29843	35142	44611	59021

Поскольку мы знаем, что потребитель кофе может быть либо мужчиной, либо женщиной, то можем объединить эти показатели, чтобы получить более широкое видение потребления в целом за отдельный год или весь период наблюдения в целом.

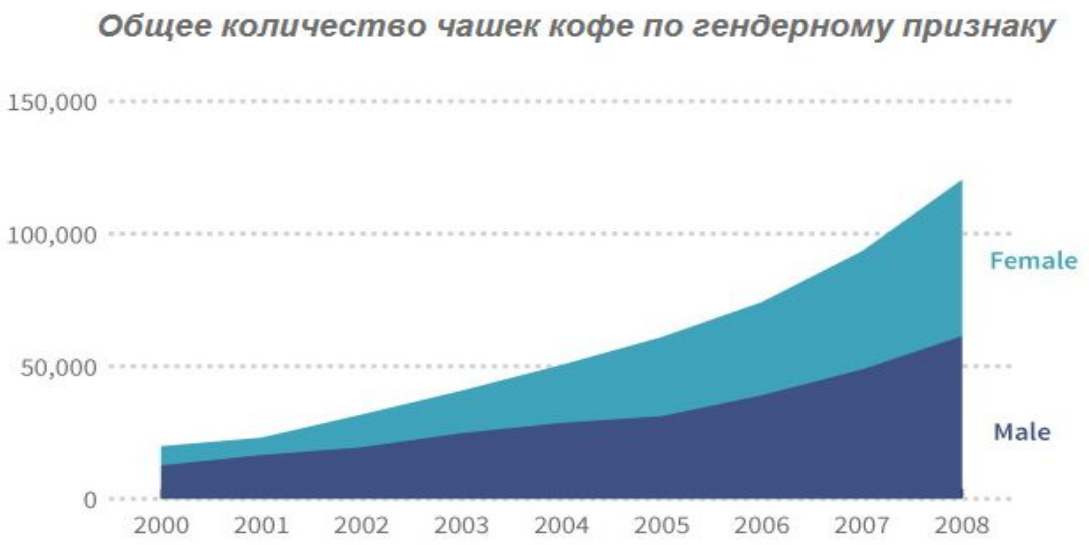
Визуализация суммируемых многорядов

Прежде всего, мы можем продемонстрировать процентное соотношение:

Анализ данных. Виды данных

Визуализация суммируемых многорядов

Кроме того, мы можем сложить сегменты суммируемого многоряда и показать целостную картину:



Анализ данных. Виды данных

Проблемы суммируемых многорядов

Сложность при работе с **суммируемыми многорядами** заключается в том, что необходимо точно знать, какие ряды совместимы друг с другом.

Год	2000	2001	2002	2003	2004
Мужчины	12534	16452	19362	24726	28567
Женщины	7261	6553	12349	16002	21873
Обыкновенный	9929	14021	17364	20035	27854
Без кофеина	6744	6833	10201	13462	17033
Мокко	3122	2151	4146	7231	5553

В этих данных нет ничего, что дало бы нам возможность объединить всю информацию. Необходимо человеческое понимание категорий данных, чтобы знать, что *мужчины + женщины = полный набор*, а также *обычный кофе + кофе без кофеина + мокко = полный набор*. Без этого знания мы не можем объединить данные или, что еще хуже, можем объединить их неправильно.

Анализ данных. Виды данных

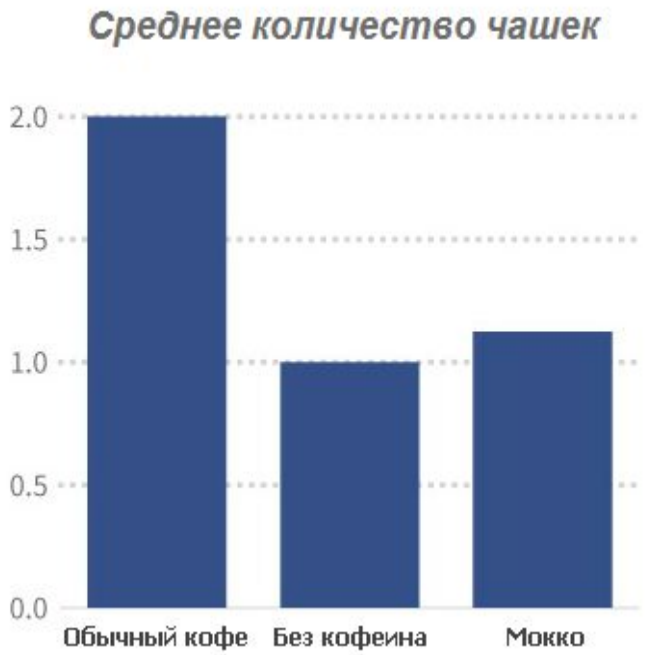
Агрегированные записи

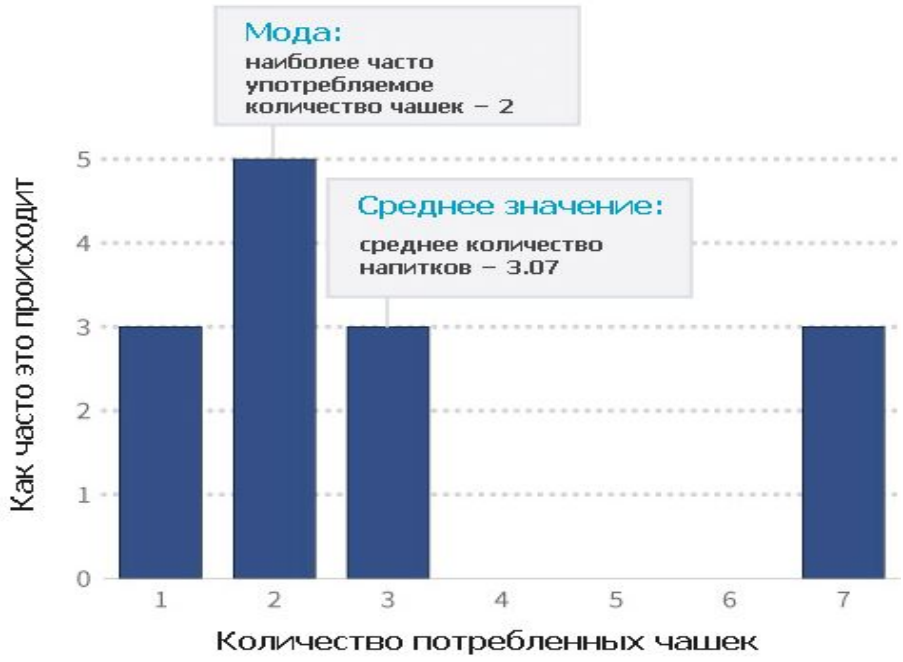
Следующая таблица с **агрегированными записями** включает колонку с категориальной информацией (пол, с двумя возможными вариантами) и промежуточные суммы для каждого типа кофе. Кроме того, в нее входят итоговые суммы для этих типов.

Имя	Пол	Обычный кофе	Кофе без кофеина	Мокко	Итого
Боб Смит	М	2	3	1	6
Джейн Доу	Ж	4	0	0	4
Дейл Купер	М	1	2	4	7
Мэри Бруер	Ж	3	1	0	4
Бетти Кона	Ж	1	0	0	1
Джон Ява	М	2	1	3	6
Билл Бин	М	3	1	0	4
Джейк Битник	М	0	0	1	1
ИТОГО	5М, 3Ж	16	8	9	33

Анализ данных. Виды данных

Визуализация результатов агрегирования





Анализ данных. Виды данных

Отдельные транзакции

Транзакционные (<сырые>) записи представляют собой данные о конкретных событиях. Здесь нет агрегации данных вокруг какого-либо параметра. Данные не накапливают во времени, они одномоментны. Но именно они и представляют наибольший интерес для аналитиков. Пример:

Отметка времени	Имя	Пол	Кофе
17:00	Боб Смит	М	Обычный
17:01	Джейн Доу	Ж	Обычный
17:02	Дейл Купер	М	Мокко
17:03	Мэри Бруер	Ж	Без кофеина
17:04	Бетти Кона	Ж	Обычный
17:05	Джон Ява	М	Обычный
17:06	Билл Бин	М	Обычный
17:07	Джейк Битник	М	Мокко
17:08	Боб Смит	М	Обычный
17:09	Джейн Доу	Ж	Обычный
17:10	Дейл Купер	М	Мокко
17:11	Мэри Бруер	Ж	Обычный
17:12	Джон Ява	М	Без кофеина
17:13	Бил Бин	М	Обычный

Анализ данных. Виды данных

Основные источники данных – подведем итог

- фактоиды
- ряды
- временные ряды
- многоряды
- суммируемые многоряды
- агрегированные записи
- отдельные транзакции

Анализ данных. Виды данных