Санкт-Петербургский
государственный
университет
www.spbu.ru

Анализ данных
Основные принципы

Графеева Н.Г.

2018

Анализ данных. Основные принципы

Последовательность работы

Анализ данных. Основные принципы

Способы анализа данных

Главным лицом в процессе анализа является эксперт – специалист в предметной области.

Несмотря на то, что существует большое количество аналитических задач, методы их решения можно поделить на две категории:

- извлечение, агрегирование и визуализация данных
- построение и использование моделей

Анализ данных. Основные принципы

Общая схема анализа

Анализ данных. Основные принципы

Визуализация данных

Эксперт формулирует запросы к имеющимся данным, возможно, агрегирует результаты запросов и отображает в виде:

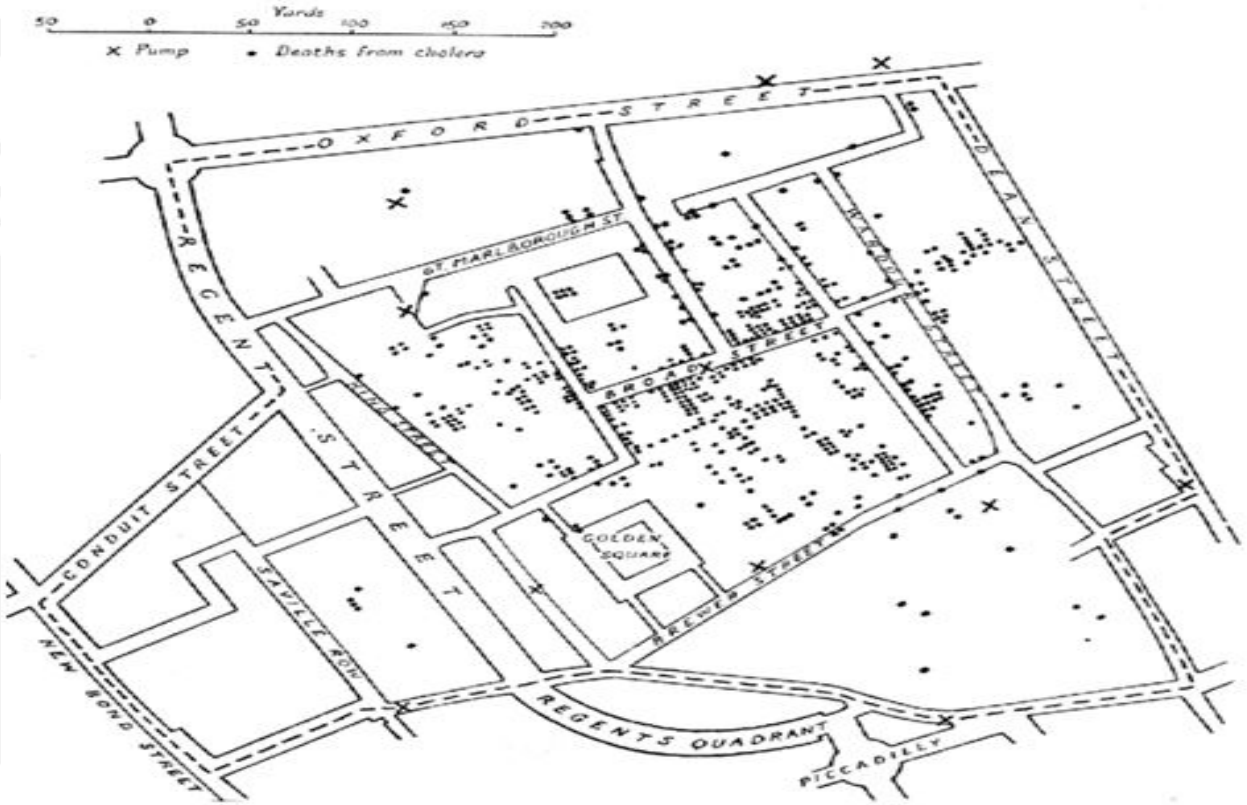
графиков, диаграмм, гистограмм,таблиц, схем, карт и т.п.

Как ни странно – эти простые методы иногда неплохо работают.

Далее – 2 примера (локализация очага холеры в Лондоне в 1854 г. и визуализация причин смертельных случаев в Крымской войне 1855 г.)

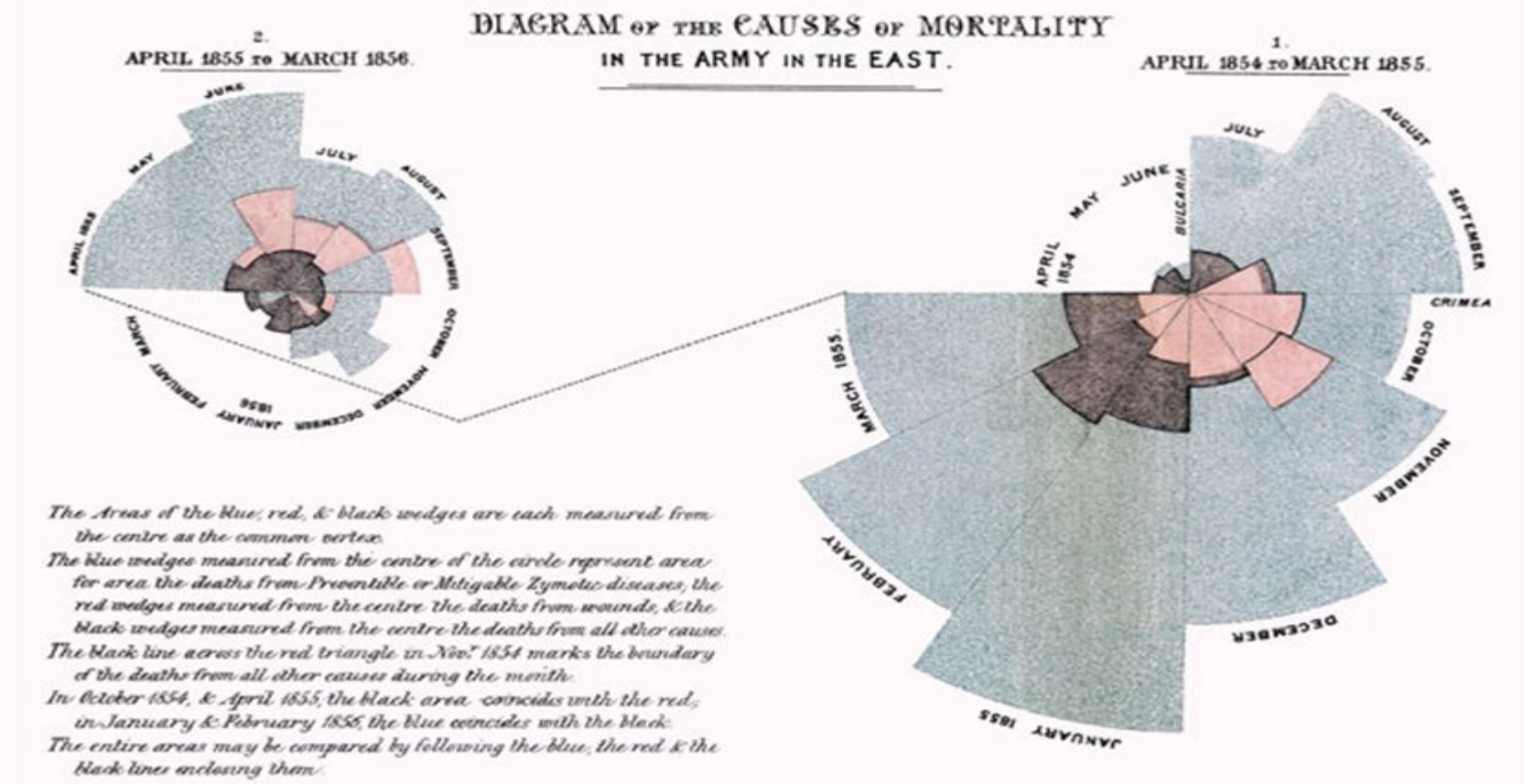
Анализ данных. Основные принципы

Пример (карта распространения холеры в Лондоне, составленная в 1854 году эпидемиологом Джоном Сноу)



Анализ данных. Основные принципы

Пример (диаграмма, составленная медсестрой Флоренс Найнтингейл в 1855 году)



Анализ данных. Основные принципы

Достоинства и недостатки визуализации

Достоинства:

- Простота создания
- Внятная интерпретация результатов

Недостатки:

- Нет возможности для анализа сложных закономерностей
- Зависит от профессионализма эксперта
- Нет возможности для тиражирования

Анализ данных. Основные принципы

Построение моделей

- Построение моделей является достаточно универсальным способом для решения многих аналитических задач. Этот способ дает возможность прогнозировать, разбивать на группы и т. п. Но самое главное – он позволяет в дальнейшем тиражировать модели для аналогичных случаев.

Анализ данных. Основные принципы

Методика извлечения знаний

Несмотря на большое количество разнообразных бизнес-задач почти все они решаются по единой методике. Эта методика называется **Knowledge Discovery in Databases**. Она описывает не конкретный алгоритм или математический аппарат, а последовательность действий, которую необходимо выполнить для построения модели (извлечения знания). Данная методика не зависит от предметной области, это набор атомарных операций, комбинируя которые можно получить нужное решение.

Анализ данных. Основные принципы

Knowledge Discovery in Databases

Анализ данных. Основные принципы

KDD – выборка данных

Первым шагом в анализе является получение исходной выборки. На основе этих данных и строятся модели. На этом шаге необходимо активное участие эксперта для выдвижения гипотез и отбора факторов, влияющих на анализируемый процесс. Желательно, чтобы данные были уже собраны и консолидированы. Крайне необходимо наличие удобных механизмов подготовки выборок. В качестве источника рекомендуется использовать специализированное хранилище данных, агрегирующее всю необходимую для анализа информацию.

Анализ данных. Основные принципы

KDD – очистка данных

Реальные данные для анализа редко бывают хорошего качества. Необходимость предварительной обработки при анализе данных возникает независимо от того, какие технологии и алгоритмы используются. Более того, эта задача может представлять самостоятельную ценность в областях, не имеющих непосредственного отношения к анализу данных. К задачам очистки относятся:

- Заполнение пропусков и редактирование аномалий
- Сглаживание, очистка от шумов
- Редактирование дубликатов и противоречий
- Устранение незначащих факторов и прочее…

Анализ данных. Основные принципы

KDD – трансформация данных

Трансформация данных – последний этап перед, собственно, анализом. Различные алгоритмы анализа требуют специальным образом подготовленные данные, например, для прогнозирования необходимо преобразовать временной ряд при помощи скользящего окна. Задачи трансформации данных:

- Нормализация данных
- Агрегирование данных по скользящему окну
- Приведение типов
- Выделение временных интервалов
- Преобразование непрерывных значений в дискретные и наоборот
- Сортировка, группировка, агрегация и прочее…

Анализ данных. Основные принципы

KDD – Data Mining

Data Mining – это процесс обнаружения в «сырых» данных, ранее неизвестных и нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Информация, найденная в процессе применения методов Data Mining, должна быть нетривиальной и ранее неизвестной, например, средние продажи не являются таковыми. Знания должны описывать новые связи между свойствами, предсказывать значения одних признаков на основе других.

Анализ данных. Основные принципы

KDD – интерпретация

В случае, когда извлеченные знания непрозрачны для пользователя, должны существовать методы постобработки, позволяющие привести их к интерпретируемому виду. Для оценки качества полученной модели нужно использовать как формальные методы оценки (всевозможные метрики), так и знания эксперта. Полученные модели являются по сути формализованными знаниями эксперта, поэтому их можно и нужно тиражировать.

Анализ данных. Основные принципы

Достоинства и недостатки моделей

Достоинства:

- Возможность тиражирования знаний
- Обработка огромных объемов данных
- Обнаружение нетривиальных закономерностей
- Формализация процесса принятия решений

Недостатки:

- Строгие требования к качеству и количеству данных
- Неспособность анализировать нестандартные случаи
- Высокие требования к знаниям эксперта.

Анализ данных. Основные принципы

Аналитическая система

Наиболее оптимальной с точки зрения гибкости, возможностей и простоты использования является аналитическая система состоящая из хранилища данных, механизмов визуализации и методов построения моделей. Подобная система позволяет комбинировать подходы к анализу данных. На стыке использования различных методов анализа получаются наиболее интересные результаты.

Анализ данных. Основные принципы

Решаемые бизнес-задачи

Подавляющее большинство бизнес-задач сводится к комбинированию описанных методов. Фактически, ранее были описаны базовые блоки, из которых собирается практически любое бизнес-решение:

- План-факторный анализ – визуализация данных
- Прогнозирование – задача регрессии.
- Управление рисками – регрессия, кластеризация и классификация.
- Стимулирование спроса – кластеризация, ассоциация
- Оценка эластичности спроса – регрессия.
- Выявление предпочтений клиентов – последовательность, кластеризация…

Анализ данных. Основные принципы