Санкт-Петербургский
государственный
университет
www.spbu.ru

Анализ данных

Подготовка данных

Графеева Н.Г.

2018

Анализ данных.Подготовка данных

Основные этапы подготовки данных

- Загрузка данных в хранилища
- Разделение данных
- Приведение данных к одинаковым единицам измерения
- Преобразование к унифицированной лексике
- Объединение данных из разных источников
- Соединение данных из разных источников
- Заполнение отсутствующих значений
- Очистка данных (устранение дубликатов, проверка шаблонов, контроль диапазонов)

Анализ данных. Подготовка данных

Загрузка данных в хранилища

Как правило, в системах хранения данных существуют специальные утилиты, ориентированные на загрузку данных из внешних источников. Однако, даже на этом, казалось бы простейшем, этапе исследователя могут ожидать многочисленные сюрпризы: например, нечитаемые символы, типы данных не соответствующие обещанным спецификациям и т.п. Рекомендации:

- Вычистить из исходных файлов все нечитаемые символы;
- Записать все исходные данные как текстовые поля (с типами разбираться потом после загрузки в хранилище).
- Саму загрузку (если данных действительно много) проводить непосредственно на сервере, где расположено хранилище.

Анализ данных. Подготовка данных

Разделение данных

Простой пример задачи, с которой сталкиваются многие люди, – это разделение имен и фамилий (или адресов). У вас может быть база данных, где имена и фамилии прописаны в одной ячейке, а вам нужно их отделить друг от друга. Или у вас уже могут быть отдельные ячейки для имен и фамилий, но в некоторых случаях имена с фамилиями все равно записаны вместе. Например:

Все в одной ячейке
Полное имя
Keith Pallard
Fumi Takano
Rhonda Johnson
Warren Andersen
Juan Tyler
Cicely Pope

В двух ячейках, но не все имена записаны правильно	
Имя	**Фамилия**
	Keith Pallard
Fumi Takano	
Rhonda	Johnson
Warren Andersen	
Juan	Tyler
Cicely	Pope

Анализ данных. Подготовка данных

Данные, также требующие разделения

Инициалы в середине имени	Martina C. Daniels
Обозначения профессии	Lloyd Carson DVM (доктор ветеринарной медицины)
Двойные фамилии	Lora de Carlo
Звания, титулы	Rev (преподобный) Herman Phillips
Окончания	Jimmy Walford III
Фамилии, написанные через дефис	Tori Baker-Andersen
Фамилия, написанная перед именем	Kincaid Jr, Paul
Двойные имена	Ray Anne Lipscomb
Звания и титулы с окончаниями	Rev Rhonda-Lee St. Andrews-Fernandez, DD, MSW
Неверно включенная ячейка	Murray Wilkins 993 E Plymouth Blvd
Отсутствует имя	O'Connor
Отсутствует фамилия	Tanya
Вообще нет имени	
Черт знает что	JJ
Имя, принадлежащее не человеку	North City Garden Supply

Анализ данных. Подготовка данных

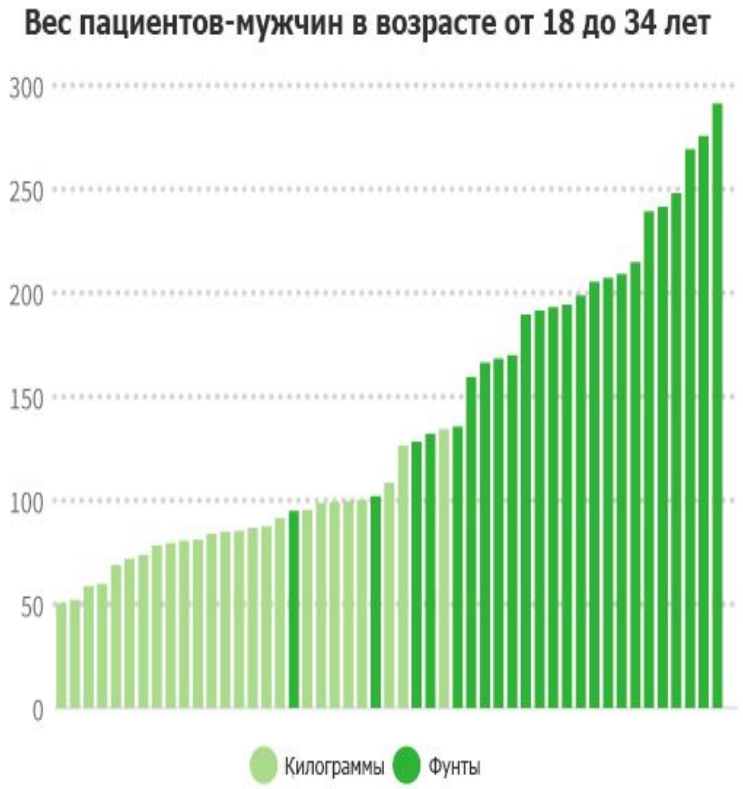
Пример (разнообразие имен из реального хранилища)

Все должности; без уволенных сотрудников; принятые на работу с 14/01/99 по 12/05/18				
Фамилия	Имя	Отчество	Должность	Дата поступления
Абдибаитова	Света		стюард	01/07/2006
Абдикеримов	Нурлан	Жетимович	стюард	10/11/2008
Абибилла Кызы	Умитай		Посудомойщик на Производс	21/09/2015
Абибилла Кызы	Бирлкан		Подсобный рабочий на Прои	01/09/2015
Абибилла Кызы	Калипа		Подсобный рабочий на Прои	31/03/2017
Аблитова	Азадакан	Маманазаровна	стюард	09/09/2008
Абраамова	Ирина		посудомойщик	10/06/2010
Адыкова	Надя		стюард	01/07/2006
Азаренкова	Ольга	Михайловна	Кондитер/час	17/07/2006
Азизова	Назгул	Жапаровна	Посудомойщик на Производс	27/11/2015
Айлчи Улы	Темирлан		Подсобный рабочий на Прои	29/08/2016
Акматова	Аинура		Подсобный рабочий на Прои	18/07/2017
Акого	Ж.		посудомойщик	06/09/2010

Анализ данных. Подготовка данных

Преобразование данных к одинаковым единицам измерения

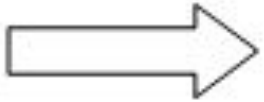
Еще один важный момент при подготовке данных − проверить, чтобы все данные в одной колонке были представлены в одинаковых единицах. Например, у вас могут быть медицинские данные из разных стран, где в одних странах вес измерен в фунтах, а в других − в килограммах. Важно конвертировать все числа или в килограммы, или в фунты, чтобы они измерялись по одной шкале, иначе их нельзя будет сравнивать и агрегировать, и какую бы вы не делали визуализацию таких необработанных данных, она будет выглядеть довольно странно. Например:



Анализ данных. Подготовка данных

Пример преобразование данных к одинаковым единицам измерения

Пациент	Вес
Джон	130 lb
Михаил	81 кг
Саймон	150 lb
Антон	75 кг
Денис	65 кг

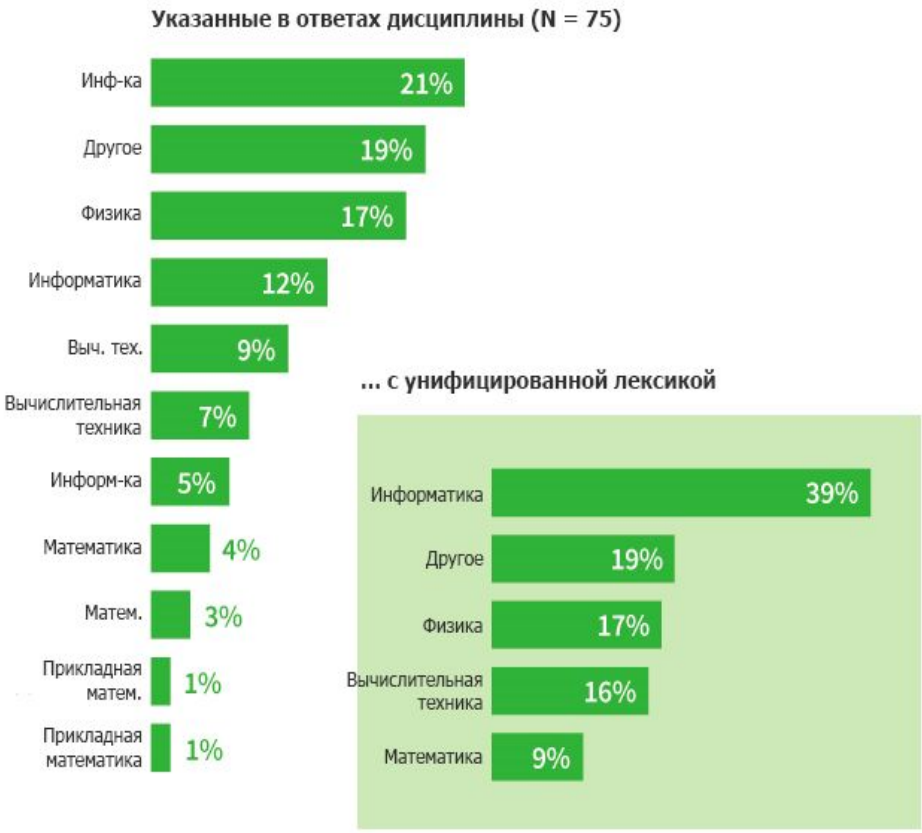


Пациент	Вес (кг)
Джон	59 кг
Михаил	81 кг
Саймон	68 кг
Антон	75 кг
Денис	65 кг

Анализ данных. Подготовка данных

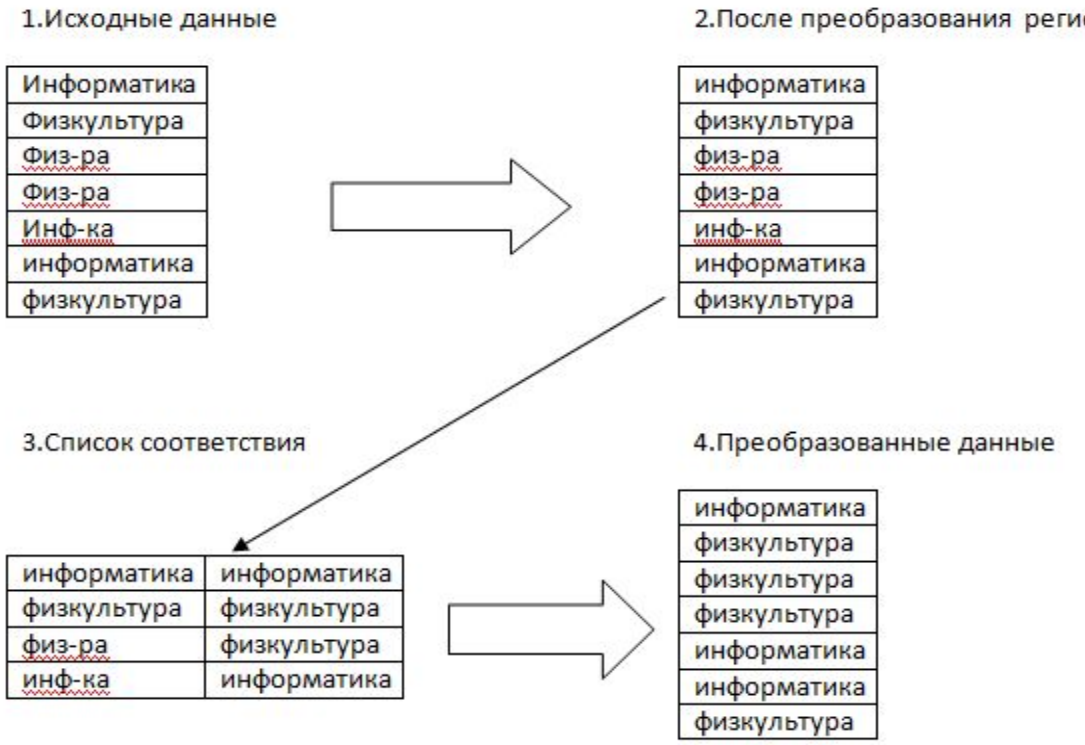
Преобразование к унифицированной лексике

Одной из самых трудоемких задач при очистке данных является работа с несовместимой информацией. Например, одно из текстовых полей в исходных данных содержит сведения о профильной дисциплине студентов. Один студент может ответить «Инф-ка», другой − «Информатика», а третий − «Информ-ка». Даже если вы знаете, что все эти ответы обозначают одну и ту же дисциплину, они крайне ограничат возможности для агрегирования и могут привести к неадекватным результатам. Необходимо преобразовывать данные к унифицированной лексике.



Анализ данных. Подготовка данных

Пример преобразования к унифицированной лексике



Анализ данных. Подготовка данных

Объединение данных из разных источников

1 подъезд

Ответственный жилец	Номер квартиры
Семен	11
Анна	12
Евгений	13

2 подъезд

Жилец	Номер квартиры	Общее количество жильцов в квартире
Сидор	14	4
Аделаида	15	2
Иннокентий	16	4

Анализ данных. Подготовка данных

Объединение данных из разных источников. Вариант 1

1.Исходные данные по 2 подъезду

Жилец	Номер квартиры	Общее количество жильцов в квартире
Сидор	14	4
Аделаида	15	2
Иннокентий	16	4

2.Преобразованные данные по 2 подъезду

Ответственный жилец	Номер квартиры
Сидор	14
Аделаида	15
Иннокентий	16

3.Обьединение данных по двум подъездам

Ответственный жилец	Номер квартиры
Семен	11
Анна	12
Евгений	13

union

Ответственный жилец	Номер квартиры
Сидор	14
Аделаида	15
Иннокентий	16

=

Ответственный жилец	Номер квартиры
Семен	11
Анна	12
Евгений	13
Сидор	14
Аделаида	15
Иннокентий	16

Анализ данных. Подготовка данных

Объединение данных из разных источников. Вариант 2

1.Исходные данные по 1 подъезду

Ответственный жилец	Номер квартиры
Семен	11
Анна	12
Евгений	13

2.Преобразованные данные по 1 подъезду

Ответственный жилец	Номер квартиры	Общее количество жильцов в квартире
Семен	11	
Анна	12	
Евгений	13	

3.Объединение данных по двум подъездам

Ответственный жилец	Номер квартиры	Общее количество жильцов в квартире
Семен	11	
Анна	12	
Евгений	13	

union

Ответственный Жилец	Номер квартиры	Общее количество жильцов в квартире
Сидор	14	4
Аделаида	15	2
Иннокентий	16	4

=

Ответственный жилец	Номер квартиры	Общее количество жильцов в квартире
Семен	11	
Анна	12	
Евгений	13	
Сидор	14	4
Аделаида	15	2
Иннокентий	16	4

Анализ данных. Подготовка данных

Соединение данных из разных источников

- Первая проблема – соответствие полей. Так же, как это было в задаче объединения данных из разных источников, необходимо исследовать соответствие полей и преобразовать названия к единому стилю.
- Вторая проблема – преобразование данных в различных источниках к единым шкалам, единицам измерения и унифицированной лексике.
- Третья проблема – идентификация данных, относящихся к одному и тому же объекту (например, выявление данных, про одного и того же покупателя в разных супермаркетах).
- И наконец, сами источники данных могут быть представлены в виде структур различных форматов (таблицы, JSON, XML и т.п.).

Анализ данных. Подготовка данных

Пример соединения данных из разных источников

1.Исходные данные

Данные из спортивного клуба

Имя	Фамилия	Дата рождения	E-mail	телефон	паспорт	Вид спорта
Никита	Семенов	08.02.1998	N.Semenov@mail.ru	8(922)468-2929	4004 271492	плавание
Юра	Алексеев	03.01.1999	Y.Alexeev@mail.ru	8(931)852-9582	3003 262899	волейбол

Данные из супермаркета

Имя покупателя	Фамилия покупателя	e-mail	телефон	Категория
Ник	Семенов	N.Semenov@mail.ru	8(922)468-2929	студент
Юрий	Алексеев	Y.Alexeev@mail.ru	8(931)852-9582	служащий

2.Результат соединения

Имя	Фамилия	Дата рождения	E-mail	телефон	паспорт	Вид спорта	Категория
Никита	Семенов	08.02.1998	N.Semenov@mail.ru	8(922)468-2929	4004 271492	плавание	студент
Юрий	Алексеев	03.01.1999	Y.Alexeev@mail.ru	8(931)852-9582	3003 262899	волейбол	служащий

Анализ данных. Подготовка данных

Заполнение отсутствующих численных значений

Одна из самых раздражающих проблем при работе с данными – пустые или не полностью заполненные поля. Если данные просто не были собраны, возможно, вы сможете вернуться к источнику и заполнить пробелы, но возможно, что у вас больше не будет доступа к этому источнику. Например, это показания датчиков, и никаких других данных просто не будет. Есть два подхода при работе с такими данными:

- Выделение таких полей специальными значениями (и исключение их из дальнейшего анализа).
- Аппроксимация пропущенных значений на основе исторических данных.

Анализ данных. Подготовка данных

Аппроксимация пропущенных значений

В большинстве случаев (особенно во временных рядах) аппроксимация пропущенных значений осуществляется за счет определения ближайших соседей и вычисления их среднего значения. Однако в некоторых случаях приходится пользоваться значительно менее стандартными алгоритмами. Например, при прохождении маршрута были потеряны сведения о времени прохождения нескольких последовательных остановок. Надо восстановить это время на основе исторических данных и временам, зафиксированным до потери и после.

Анализ данных. Подготовка данных

Пример (пропущенные значения)

Остановка	Время прибытия	Номер остановки
Университет	16:00	1
Общежития	?	2
23 квартал	?	3
Старый Петергоф	?	4
Ж.Д. переезд	?	5
Часовой завод	?	6
Фонтаны	?	7
Новый Петергоф (вокзал)	16:45	8

Анализ данных. Подготовка данных

Очистка данных

Как правило, очистка данных может быть сведена к выполнению следующих работ:

- проверка сочетания полей
- сравнение с образцом/регулярные выражения
- устранение дубликатов
- контроль диапазонов

Анализ данных. Подготовка данных

Сочетание полей

- Для проверки данных можно также использовать сочетание полей. Иногда это действительно необходимо, потому что нужно взглянуть на все поля в записи, чтобы определить одно или несколько неправильных. Представьте, что вы получили данные медицинского обследования пациентов в больнице и отслеживаете принимаемые ежедневно лекарства, используя три отдельных поля для данных: название лекарства, назначенная доза и единица измерения дозы препарата. То есть, если в наборе данных указано «Аспирин, 500, мг», значит, что пациент ежедневно принимал 500 мг аспирина. Теперь представьте, что вы получили запись “Морфин, 200, фунт”. Какой будет ваша реакция? Необходимо предусмотреть правила целостности, которые не допустят использование таких данных.

Анализ данных. Подготовка данных

Сравнение с образцом/Регулярные выражения

- Другой тип проверки данных, включает в себя сравнение с образцом. Такой вид проверки можно использовать, например, чтобы удостовериться, что все записи в поле – электронные адреса. Для этого используются, так называемые, “регулярные выражения” (regular expressions – regex) с помощью которых вы задаете шаблон выражения. Способ, которым вы задаете шаблон варьируется от используемого программного обеспечения, но на сегодняшний день присутствует практически в любых системах. Примеры регулярных выражений:

@.ru

DDD.DD

Анализ данных. Подготовка данных

Устранение дубликатов

- Одна из проблем, решаемая на этапе очистки данных, это устранение дубликатов. Дубликаты могут появляться в исходных данных по причине разного рода технических сбоев и могут быть причиной получения неверных результатов при последующем агрегировании данных. Пример:

Код транспорта	Название транспорта
1	автобус
2	трамвай
2	трамвай
3	троллейбус
4	метро

Анализ данных. Подготовка данных

Контроль диапазонов

Контроль диапазонов – это на первый взгляд очень простая процедура, которую мы используем в числовых полях, чтобы увидеть, находятся ли какие-либо значения в этом наборе данных выше или ниже крайних допустимых значений для этой переменной. Возьмем для примера оценки за домашнее задание. Представьте, что вы – преподаватель и внесли первую партию оценок за домашние работы за семестр. Вы хотите убедиться, что все внесено верно, поэтому открываете базу данных и сортируете ее по колонке с оценками за домашнюю работу, оцененную по шкале от 0 до 100. Вот как выглядят первые строки:

Анализ данных. Подготовка данных

Пример (контроль диапазонов)

Вот как выглядят первые строки отсортированной таблицы с отметками·

ИН студента	Оценка
679372531	980
673540288	99
674082892	97
673923590	96

Вот как выглядят последние строки таблицы:

ИН студента	Оценка
674472019	78
679029425	75
671822390	74
671278927	9

Анализ данных. Подготовка данных

Контроль диапазонов

В примере с оценками визуального анализа вполне достаточно для обнаружения и последующего исправления <криминальных> случаев. Как быть, когда данных значительно больше и они не так очевидны по содержанию? Как обнаружить редкие, но тем не менее существующие, так называемые, <выбросы данных>? И тут оказывается, что все не так просто, а в математической статистике для этого есть подходящие понятия **дисперсии**, **стандартного отклонения** и **неравенство Чебышева**.

Анализ данных. Подготовка данных

Дисперсия

Дисперсия выборки – среднее арифметическое квадратов отклонений значений выборки от выборочного среднего. Вычисляется по формуле:

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Анализ данных. Подготовка данных

Пример (вычисление дисперсии)

Имеется выборка из четырех значений: 2, 3, 6, 9

Сначала находим среднее:

$$n = 4$$

$$\bar{x} = \frac{2+3+6+9}{4} = 5$$

Теперь по формуле вычисляем дисперсию:

$$s^2 = \frac{(2-5)^2 + (3-5)^2 + (6-5)^2 + (9-5)^2}{4-1} = 10$$

Анализ данных. Подготовка данных

Стандартное отклонение

Стандартное отклонение вычисляется как корень квадратный из дисперсии:

$$s = \sqrt{s^2}$$

Стандартное отклонение имеет исключительную важность для
описания распределения данных.

Анализ данных. Подготовка данных

Неравенство Чебышева

Для интерпретации стандартного отклонения используют **неравенство Чебышева**. Оно имеет следующую трактовку:

В любой совокупности данных доля значений, попадающих в интервал

$$\overline{x} \pm ks$$

будет равна, по крайней мере,

$$1 - \frac{1}{k^2}$$

где k - любое число, большее 1.

Анализ данных. Подготовка данных

Интерпретация стандартного отклонения

Можно утверждать, что интервал с границами

$$\bar{x} \pm 2s$$

содержит, по крайней мере, 3/4 всех данных (75%).
Интервал с границами

$$\bar{x} \pm 3s$$

содержит, по крайней мере, 8/9 всех данных (89,9%).
Значения, которые не попадают в интервал, можно считать выбросами.

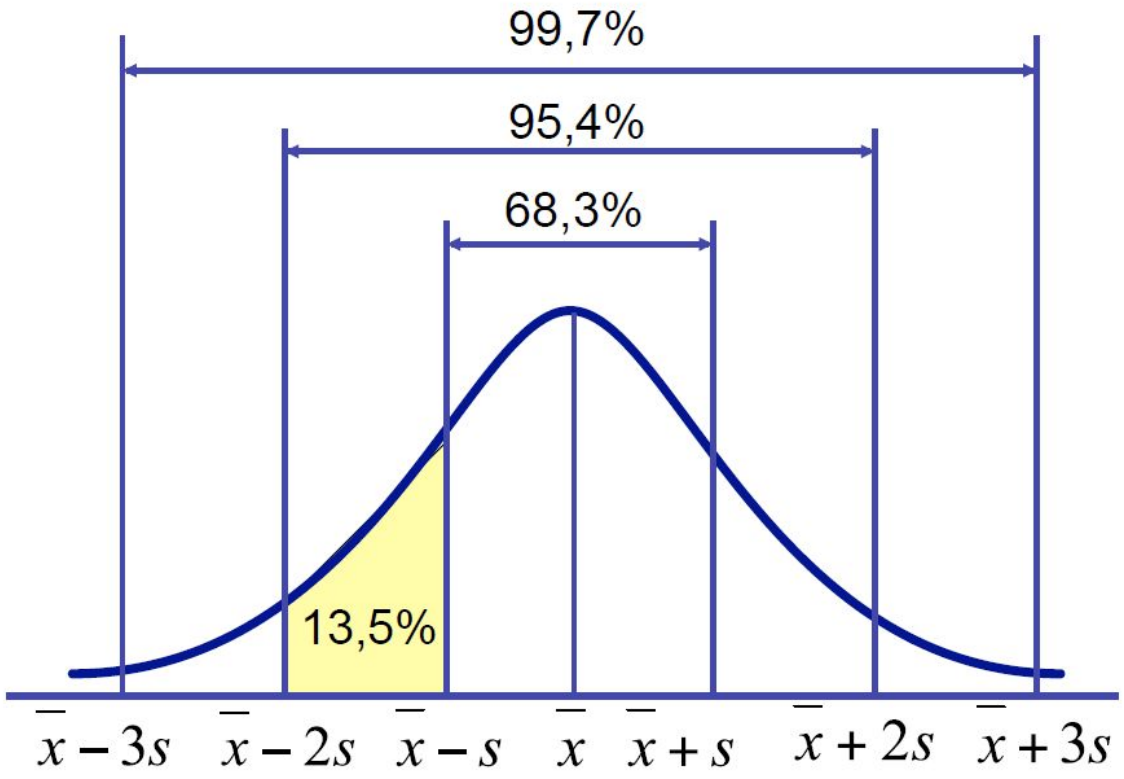
Анализ данных. Подготовка данных

Интерпретация стандартного отклонения

В математической статистике доказывают что….

Анализ данных. Подготовка данных

Для нормального распределения данных…



Анализ данных. Подготовка данных

Контроль диапазонов (итоги)

- Для определения выбросов используется понятие стандартного отклонения. Как правило – достаточно коэффициента **k** равного 3. Что делать с пропущенными значениями после исключения выбросов? Аппроксимировать их как средние или (для временных рядов) с помощью ближайших соседей (например, предыдущее и последующее значения).

Анализ данных. Подготовка данных

Основные этапы подготовки данных – подведем итог

- Загрузка данных в хранилища
- Разделение данных
- Приведение данных к одинаковым единицам измерения
- Преобразование к унифицированной лексике
- Объединение данных из разных источников
- Соединение данных из разных источников
- Заполнение отсутствующих значений
- Очистка данных (контроль диапазонов,сравнение с образцом/регулярные выражения, сочетание полей, устранение дубликатов)

Анализ данных. Подготовка данных

Задание 3

Рассчитайте дисперсию, стандартное отклонение, а затем определите выбросы в одном из своих dataset (желательно для данных с нормальным распределением). Аппроксимируйте значения после удаления выбросов. Визуализируйте результат (что было и что стало).

Примечание: Срок сдачи: 2 недели с момента выдачи. Задание отправлять по адресу: N.Grafeeva@spbu.ru.

Topic: DataMining_2018_job3

Анализ данных. Подготовка данных