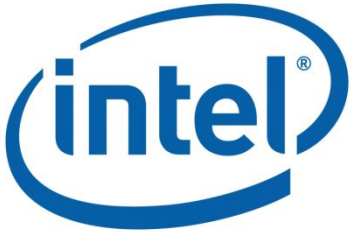
Лабораторная работа № 1 Первичный анализ наборов данных



Наборы данных

- Набор данных - объекты и признаки
- Признаки - числовые и категориальные
- Количество объектов как правило значительно больше количества признаков
- Данные чаще всего представляют в виде матриц (таблиц)

Виды наборов данных

1. Объект-Признак: каждая строчка - объект, каждый столбец - некоторый признак.
2. Сенсорные данные(временные ряды): каждый столбец - некоторый сенсор, каждая строчка - показатели сенсоров на некоторой временной отметке
3. Изображения: каждый пиксель закодирован некоторым образом (RGB, YCbCr)
4. Логи (журналы событий): каждая строчка - это событие, представленное в формализованном виде
5. Документы: неструктурированный набор данных, тексты

Пример: Turkey Student Evaluation*

- Набор данных содержит ответы студентов на вопросы о качестве преподавания предметов
- Каждый вопрос оценивается баллами от 1 до 5
- 28 вопросов о качестве преподавания по пройденному предмету
- 3 преподавателя, 13 предметов
- 5820 объектов (записей)

*http://archive.ics.uci.edu/ml/datasets/Turkiye+Student+Evaluation

instr	class	nb.repeat	attendance	difficulty	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
1	2	1	3	5	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1	2	1	3	4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1	2	1	0	1	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
1	2	1	3	5	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1	2	1	3	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
1	2	1	3	4	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	2	2	1	5	3	3	2	2	5	3	3	3	5	5	4	4	3	4	4	4	4	4	4	4
1	2	1	2	4	1	1	4	2	3	3	2	2	2	2	3	2	4	3	3	3	5	2	3	3
1	7	3	0	4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1	7	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	7	3	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1	7	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Пример: Turkey Student Evaluation

- Как можно привести данные к единообразному виду?
- Какие есть инструменты для работы с данными?
- Какие простые метрики можно использовать для работы с данными?
- Как можно очистить данные от ненужных/мешающих элементов?
- Как работать с конкретными данными?

Трансформация данных

- Дискретизация: перевод числовых данных в категориальные
- Бинаризация: трансформация одного категориального признака в несколько бинарных
- Работа с текстом: Latent Semantic Analysis (LSA)
- Временные ряды: symbolic aggregate approximation(SAX), вейвлет-преобразование, Фурье преобразование и др.

номинальный признак	малое предприятие	среднее предприятие	крупное предприятие
числовой признак	1	2	3

Объект	**Признак (ном)**	**Признак (число, вар.1)**	**Признак 1 (число, вар.2)**	**Признак 2 (число, вар.2)**	**Признак 3 (число, вар.2)**
Предприятие 1	малое	1	1	0	0
Предприятие 2	малое	1	1	0	0
Предприятие 3	среднее	2	0	1	0
Предприятие 4	малое	1	1	0	0
Предприятие 5	крупное	3	0	0	1
Предприятие 6	крупное	3	0	0	1
Предприятие 7	среднее	2	0	1	0
Предприятие 8	малое	1	1	0	0
Предприятие 9	крупное	3	0	0	1

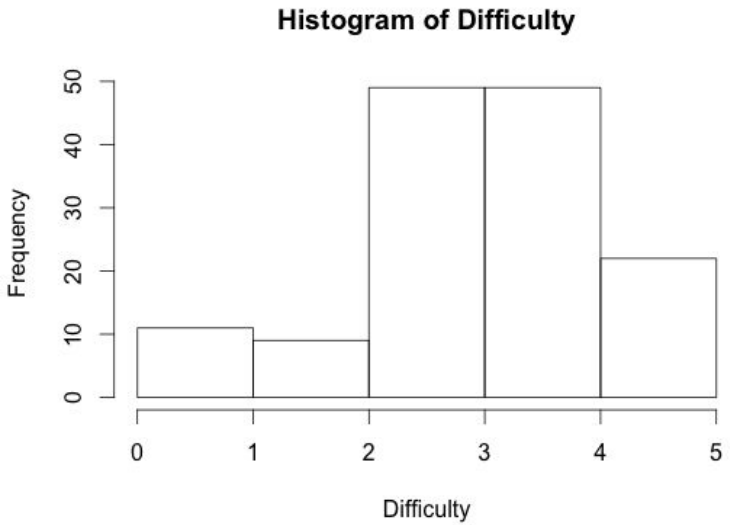
Описательные статистики

- Минимум и максимум
- Среднее значение
- Характеристики разброса
- Дисперсия
- Стандартное отклонение
- Интервал изменения
- Медиана и квантили
- Гистограмма частот
- Матрица ковариаций и корреляций (оценка связи между признаками)
- Коэффициенты асимметрии, эксцесса, высшие моменты

Turkey Student Evaluation

Сложность предмета #2 (преподаватель #1):

- Среднее значение - 3.
- Стандартное отклонен
  - 1.08
- Минимум - 1, максиму



	diff	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28
diff	1	0.05	0.07	0.07	0.06	0.06	0.05	0.05	0.05	0.06	0.04	0.06	0.04	0.08	0.09	0.09	0.05	0.12	0.07	0.08	0.09	0.1	0.1	0.08	0.07	0.1	0.06	0.06	0.09
Q1	0.05	1	0.87	0.77	0.85	0.8	0.77	0.79	0.79	0.73	0.8	0.72	0.76	0.72	0.7	0.7	0.74	0.61	0.71	0.7	0.69	0.67	0.67	0.73	0.73	0.67	0.7	0.71	0.66
Q2	0.07	0.87	1	0.85	0.87	0.86	0.83	0.84	0.83	0.8	0.85	0.79	0.8	0.8	0.79	0.79	0.81	0.72	0.79	0.79	0.78	0.76	0.77	0.8	0.8	0.77	0.78	0.77	0.75
Q3	0.07	0.77	0.85	1	0.83	0.84	0.82	0.82	0.81	0.8	0.83	0.81	0.78	0.81	0.81	0.8	0.79	0.77	0.8	0.8	0.8	0.79	0.79	0.8	0.79	0.79	0.8	0.77	0.78
Q4	0.06	0.85	0.87	0.83	1	0.87	0.84	0.84	0.82	0.78	0.84	0.77	0.79	0.78	0.77	0.78	0.79	0.7	0.77	0.77	0.76	0.75	0.75	0.79	0.79	0.75	0.77	0.76	0.74
Q5	0.06	0.8	0.86	0.84	0.87	1	0.88	0.89	0.88	0.81	0.88	0.81	0.82	0.83	0.81	0.81	0.84	0.73	0.82	0.81	0.79	0.78	0.78	0.83	0.83	0.78	0.8	0.79	0.77
Q6	0.05	0.77	0.83	0.82	0.84	0.88	1	0.89	0.86	0.8	0.87	0.8	0.81	0.81	0.8	0.8	0.82	0.72	0.78	0.79	0.78	0.77	0.77	0.8	0.8	0.77	0.79	0.78	0.76
Q7	0.05	0.79	0.84	0.82	0.84	0.89	0.89	1	0.9	0.82	0.89	0.81	0.83	0.81	0.79	0.79	0.82	0.7	0.79	0.79	0.78	0.76	0.76	0.82	0.82	0.77	0.8	0.79	0.75
Q8	0.05	0.79	0.83	0.81	0.82	0.88	0.86	0.9	1	0.83	0.89	0.81	0.84	0.79	0.78	0.77	0.82	0.7	0.79	0.78	0.77	0.75	0.75	0.81	0.82	0.76	0.79	0.79	0.73
Q9	0.06	0.73	0.8	0.8	0.78	0.81	0.8	0.82	0.83	1	0.87	0.83	0.81	0.79	0.79	0.79	0.8	0.75	0.79	0.79	0.78	0.78	0.78	0.79	0.78	0.78	0.78	0.76	0.76
Q10	0.04	0.8	0.85	0.83	0.84	0.88	0.87	0.89	0.89	0.87	1	0.86	0.87	0.84	0.82	0.82	0.86	0.73	0.83	0.82	0.81	0.8	0.8	0.84	0.84	0.79	0.83	0.82	0.78
Q11	0.06	0.72	0.79	0.81	0.77	0.81	0.8	0.81	0.81	0.83	0.86	1	0.86	0.8	0.8	0.8	0.79	0.75	0.78	0.8	0.79	0.79	0.79	0.79	0.79	0.78	0.78	0.77	0.77
Q12	0.04	0.76	0.8	0.78	0.79	0.82	0.81	0.83	0.84	0.81	0.87	0.86	1	0.79	0.77	0.76	0.8	0.69	0.78	0.78	0.76	0.75	0.74	0.79	0.8	0.74	0.77	0.77	0.73
Q13	0.08	0.72	0.8	0.81	0.78	0.83	0.81	0.81	0.79	0.79	0.84	0.8	0.79	1	0.94	0.91	0.9	0.84	0.89	0.88	0.88	0.87	0.87	0.87	0.86	0.87	0.86	0.83	0.86
Q14	0.09	0.7	0.79	0.81	0.77	0.81	0.8	0.79	0.78	0.79	0.82	0.8	0.77	0.94	1	0.93	0.89	0.88	0.89	0.89	0.9	0.89	0.89	0.87	0.86	0.89	0.86	0.83	0.87
Q15	0.09	0.7	0.79	0.8	0.78	0.81	0.8	0.79	0.77	0.79	0.82	0.8	0.76	0.91	0.93	1	0.89	0.88	0.89	0.89	0.89	0.89	0.89	0.88	0.85	0.89	0.86	0.82	0.87
Q16	0.05	0.74	0.81	0.79	0.79	0.84	0.82	0.82	0.82	0.8	0.86	0.79	0.8	0.9	0.89	0.89	1	0.8	0.91	0.88	0.87	0.85	0.85	0.89	0.88	0.85	0.86	0.85	0.83
Q17	0.12	0.61	0.72	0.77	0.7	0.73	0.72	0.7	0.7	0.75	0.73	0.75	0.69	0.84	0.88	0.88	0.8	1	0.85	0.86	0.87	0.87	0.87	0.82	0.79	0.87	0.82	0.77	0.86
Q18	0.07	0.71	0.79	0.8	0.77	0.82	0.78	0.79	0.79	0.79	0.83	0.78	0.78	0.89	0.89	0.89	0.91	0.85	1	0.9	0.88	0.87	0.87	0.88	0.87	0.86	0.86	0.83	0.84
Q19	0.08	0.7	0.79	0.8	0.77	0.81	0.79	0.79	0.78	0.79	0.82	0.8	0.78	0.88	0.89	0.89	0.88	0.86	0.9	1	0.91	0.9	0.89	0.89	0.87	0.88	0.87	0.84	0.86
Q20	0.09	0.69	0.78	0.8	0.76	0.79	0.78	0.78	0.77	0.78	0.81	0.79	0.76	0.88	0.9	0.89	0.87	0.87	0.88	0.91	1	0.93	0.91	0.89	0.86	0.89	0.87	0.83	0.88
Q21	0.1	0.67	0.76	0.79	0.75	0.78	0.77	0.76	0.75	0.78	0.8	0.79	0.75	0.87	0.89	0.89	0.85	0.87	0.87	0.9	0.93	1	0.94	0.89	0.86	0.9	0.87	0.84	0.89
Q22	0.1	0.67	0.77	0.79	0.75	0.78	0.77	0.76	0.75	0.78	0.8	0.79	0.74	0.87	0.89	0.89	0.85	0.87	0.87	0.89	0.91	0.94	1	0.9	0.87	0.91	0.87	0.84	0.89
Q23	0.08	0.73	0.8	0.8	0.79	0.83	0.8	0.82	0.81	0.79	0.84	0.79	0.79	0.87	0.87	0.88	0.89	0.82	0.88	0.89	0.89	0.89	0.9	1	0.92	0.89	0.88	0.87	0.86
Q24	0.07	0.73	0.8	0.79	0.79	0.83	0.8	0.82	0.82	0.78	0.84	0.79	0.8	0.86	0.86	0.85	0.88	0.79	0.87	0.87	0.86	0.86	0.87	0.92	1	0.88	0.88	0.87	0.84
Q25	0.1	0.67	0.77	0.79	0.75	0.78	0.77	0.77	0.76	0.78	0.79	0.78	0.74	0.87	0.89	0.89	0.85	0.87	0.86	0.88	0.89	0.9	0.91	0.89	0.88	1	0.89	0.85	0.9
Q26	0.06	0.7	0.78	0.8	0.77	0.8	0.79	0.8	0.79	0.78	0.83	0.78	0.77	0.86	0.86	0.86	0.86	0.82	0.86	0.87	0.87	0.87	0.87	0.88	0.88	0.89	1	0.88	0.88
Q27	0.06	0.71	0.77	0.77	0.76	0.79	0.78	0.79	0.79	0.76	0.82	0.77	0.77	0.83	0.83	0.82	0.85	0.77	0.83	0.84	0.83	0.84	0.84	0.87	0.87	0.85	0.88	1	0.85
Q28	0.09	0.66	0.75	0.78	0.74	0.77	0.76	0.75	0.73	0.76	0.78	0.77	0.73	0.86	0.87	0.87	0.83	0.86	0.84	0.86	0.88	0.89	0.89	0.86	0.84	0.9	0.88	0.85	1

Инструменты анализа данных

- Intel DAAL (Data Analytics Acceleration Library)
  - C++, Java, Python версии (на 2017 год)
  - Заточенность на скорость работы алгоритмов
- Python – Sci-kit Learn, Scipy + Numpy библиотеки.
  - Большое количество алгоритмов по анализу данных
  - Удобные интерфейсы и возможность построения графиков (mathplotlib)
- Язык R. Свободно распространяемое программное обеспечение для анализа данных.
  - Большое количество алгоритмов по анализу данных (иногда в нескольких вариантах) с документацией
  - Наличие универсальной IDE (R Studio)

Пример кода

```
FileDataSource dataSource = new FileDataSource(context, datasetFileName,
DataSource.DictionaryCreationFlag.DoDictionaryFromContext,
        DataSource.NumericTableAllocationFlag.DoAllocateNumericTable);

Batch algorithm = new Batch(context, Double.class, Method.defaultDense);
NumericTable input = dataSource.getNumericTable();
algorithm.input.set(InputId.data, input);

//Задаем базу и алгоритм для работы

double data[] = {0.25, 0.5, 0.75};
HomogenNumericTable quantileOrders = new HomogenNumericTable(context, data, 3, 1);
algorithm.parameter.setQuantileOrders(quantileOrders);
result = algorithm.compute();

//Задаём параметры алгоритма и производим вычисления

NumericTable table = result.get(id);
long r = table.getNumberOfRows();
long c = table.getNumberOfColumns();
DoubleBuffer buf = DoubleBuffer.allocate((int) (r * c));
buf = table.getBlockOfRows(0, r, buf);

//Вытаскиваем результаты из алгоритма
```

Вычисление описательных статистик PAL

```
FileDataSource dataSource = new FileDataSource(context, datasetFileName,
                DataSource.DictionaryCreationFlag.DoDictionaryFromContext,
DataSource.NumericTableAllocationFlag.DoAllocateNumericTable);
dataSource.loadDataBlock(140);

Batch algorithm = new Batch(context, Double.class, Method.defaultDense);
NumericTable input = dataSource.getNumericTable();
algorithm.input.set(InputId.data, input);
result = algorithm.compute();

//Распечатка результатов
DoubleBuffer buf = getData(ResultId.sum);
System.out.println("Sum: "+ buf.get(0));
buf = getData(ResultId.mean);
System.out.println("Mean: "+ buf.get(0));
buf = getData(ResultId.standardDeviation);
System.out.println("Standard deviation: "+ buf.get(0));
buf = getData(ResultId.minimum);
System.out.println("Minimum: "+ buf.get(0));
buf = getData(ResultId.maximum);
System.out.println("Maximum: "+ buf.get(0));
context.dispose();
```

```
static DoubleBuffer getData(ResultId id)
{ //Функция получения нужных результатов
    NumericTable table = result.get(id);
    long r = table.getNumberOfRows();
    long c = table.getNumberOfColumns();
    DoubleBuffer buf = DoubleBuffer.allocate((int)
(r * c));
    buf = table.getBlockOfRows(0, r, buf);
    return buf;
}
```

Вычисление описательных статистик

```
import numpy as np
import scipy as sp

data = sp.genfromtxt(datapath, delimiter = ',')
x = data[0:140,0]
y = data[0:140,1]
print(data.shape)
print("Sum: ",np.sum(x))
print("Mean: ",np.mean(x))
print("Variance: “,np.var(x))
print("Standard deviation: ",np.std(x))
print("Minimum: ",np.min(x))
print("Maximum: ",np.max(x))
print("Quantiles: “,np.percentile(x,[25,50,75])
print(np.corrcoef(x, y)[0,1])

np.std(a,axis=None,dtype=None,out=None

,ddof=0,keepdims=False)
```

```
sum(X)
mean(X)
var(X)
sd(X)
min(X)
max(X)
quantile(X,probs = c(0.25,0.5,0.75))

cor(X,Y)[1,2]


var(x, y = NULL, na.rm = FALSE, use)
```

Аномалии в данных

- Неточности в данных связанные с неточностью или ошибкой измерительных приборов, отказом оборудования
- Ошибки при сканировании, неточности, связанные с ошибкой распознавания
- Некорректная информация, полученная от людей - опрашиваемых, испытуемых.
- Ошибки при ручном создании наборов данных

Поиск аномальных объектов

- Работа с пропущенными данными
- Избавление от несогласованности данных, подозрительно выделяющихся значений признаков, работа с выбросами
- Приведение числовых признаков к некоторому стандартному виду

Объект\признак	1	2	3	4	5	число пропусков	процент пропусков
1	1.3	9.9	6.7	3.0	2.6	0	0
2	4.1	5.7			2.9	2	40
3		9.9		3.0		3	60
4	0.9	8.6		2.1	1.8	1	20
5	0.4	8.3		1.2	1.7	1	20
6	1.5	6.7	4.8		2.5	1	20
7	0.2	8.8	4.5	3.0	2.4	0	0
8	2.1	8.0	3.0	3.8	1.4	0	0
9	1.8	7.6		3.2	2.5	1	20
10	4.5	8.0		3.3	2.2	1	20
11	2.5	9.2		3.3	3.9	1	20
12	4.5	6.4	5.3	3.0	2.5	0	0
13					2.7	4	80
14	2.8	6.1	6.4		3.8	1	20
15	3.7			3.0		3	60
16	1.6	6.4	5.0		2.1	1	20
17	0.5	9.2		3.3	2.8	1	20
18	2.8	5.2	5.0		2.7	1	20
19	2.2	6.7		2.6	2.9	1	20
20	1.8	9.0	5.0	2.2	3.0	0	0
число пропусков	2	2	11	6	2	23	
процент пропусков	10	10	55	30	10		23

Объект\признак	1	2	4	5	число пропусков	процент пропусков
1	1.3	9.9	3.0	2.6	0	0
2	4.1	5.7		2.9	1	25
4	0.9	8.6	2.1	1.8	0	0
5	0.4	8.3	1.2	1.7	0	0
6	1.5	6.7		2.5	1	25
7	0.2	8.8	3.0	2.4	0	0
8	2.1	8.0	3.8	1.4	0	0
9	1.8	7.6	3.2	2.5	0	0
10	4.5	8.0	3.3	2.2	0	0
11	2.5	9.2	3.3	3.9	0	0
12	4.5	6.4	3.0	2.5	0	0
14	2.8	6.1		3.8	1	25
16	1.6	6.4		2.1	1	25
17	0.5	9.2	3.3	2.8	0	0
18	2.8	5.2		2.7	1	25
19	2.2	6.7	2.6	2.9	0	0
20	1.8	9.0	2.2	3.0	0	0
число пропусков	2	2	6	2	5	
процент пропусков	0	0	29.4	0		7.35

instr	class	nb.repeat	attendance	difficulty	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28
1	2	1	3	5	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1	2	1	3	4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1	2	1	0	1	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
1	2	1	3	5	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1	2	1	3	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
1	2	1	3	4	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	2	2	1	5	3	3	2	2	5	3	3	3	5	5	4	4	3	4	4	4	4	4	4	4	4	5	4	5	5	4	4	4
1	2	1	2	4	1	1	4	2	3	3	2	2	2	2	3	2	4	3	3	3	5	2	3	3	3	1	3	3	1	3	3	2
1	7	3	0	4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1	7	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1
1	7	3	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1	7	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	7	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	7	2	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
1	7	1	1	3	4	4	4	4	4	4	4	3	3	3	3	4	4	3	3	1	3	2	3	3	3	3	3	3	3	3	3	3
1	7	1	3	3	1	1	5	1	5	4	5	5	1	4	5	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
1	7	1	1	4	4	4	4	4	4	4	4	4	4	4	4	2	4	4	4	4	4	4	4	3	3	4	4	4	4	4	4	4
1	7	1	0	1	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
1	7	1	2	2	3	3	4	4	4	3	4	5	3	3	4	3	4	3	4	3	4	3	4	3	2	3	3	3	4	4	4	3
1	7	1	4	4	4	5	5	4	5	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
1	7	1	2	4	2	2	4	2	3	2	3	4	3	3	4	1	4	4	4	3	4	4	4	4	3	2	3	4	4	4	4	3
1	7	1	3	4	4	4	4	4	4	4	4	4	5	5	5	4	4	5	4	4	4	4	4	4	4	4	4	4	4	4	4	4

Поиск выбросов

- Поиск выбросов с использованием квартилей:
  - Q_1 - значение признака, которое больше 25% значений из данных.
  - Q_3 - значение признака, которое больше 75% значений из данных
  - Выбросом является значение вне интервала $[X_1, X_2]$

$$X_1 = Q_1 - k * (Q_3 - Q_1) \qquad X_2 = Q_3 + k * (Q_3 - Q_1)$$

Поиск выбросов

- Поиск выбросов по распределениям признаков:
  - Все объекты, для которых выполнено неравенство, являются выбросами:

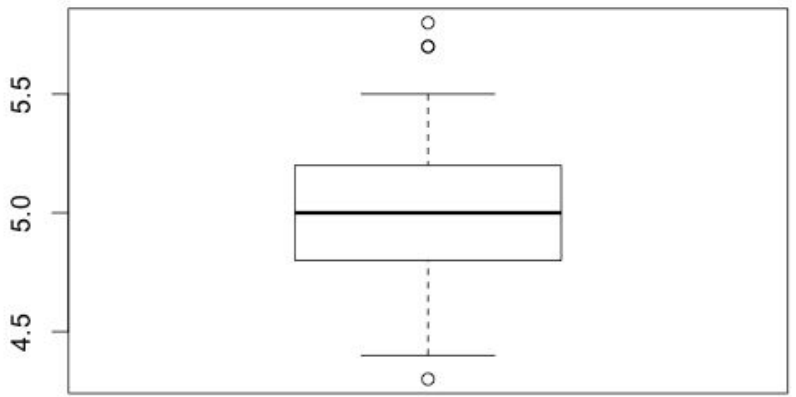
$$\sqrt{(x - \bar{x})' \Sigma_n^{-1} (x - \bar{x})} > g(n, \alpha_n)$$

где Σ – матрица ковариаций признаков.

Поиск выбросов, R

```
> L = boxplot(iris[1:50,1], range = 1)
> L$out
//функция построения диаграммы «коробка с усами»

[1] 4.3 5.8 5.7 5.7
> L$stats
     [,1]
[1,]  4.4
[2,]  4.8
[3,]  5.0
[4,]  5.2
[5,]  5.5
```



Поиск выбросов, DAAL

```
Batch alg = new Batch(context, Double.class, Method.baconDense);
NumericTable table = dataSource.getNumericTable();

//Установка параметров
alg.parameter.setInitializationMethod(InitializationMethod.baconMedian);
alg.parameter.setAlpha(0.01);
alg.input.set(InputId.data, table);

//Вычисление и получение результатов
Result result = alg.compute();
NumericTable weights = result.get(ResultId.weights);
long r = weights.getNumberOfRows();
long c = weights.getNumberOfColumns();
DoubleBuffer buf = DoubleBuffer.allocate((int) (r*c));
buf = weights.getBlockOfRows(0,r,buf);
for(int i = 0; i < r; i++)
    if(buf.get(i) == 0)
```

Стандартизация данных

- Стандартизация: $a^j = \min_i x_i^j \quad b^j = \max_i x_i^j$

  1) $z_i^j = \dfrac{2x_i^j - (b^j + a^j)}{b^j - a^j} \quad z_i^j \in [-1, 1]$

  2) $z_i^j = \dfrac{x_i^j - a^j}{b^j - a^j} \quad z_i^j \in [0, 1]$

- Нормализация: $\mu^j = \overline{x^j} \qquad \sigma_j^2 = \dfrac{1}{n-1} \sum_i (x_i^j - \overline{x^j})^2$

$$z_i^j = \frac{x_i^j - \mu^j}{\sigma^j} \qquad \overline{z^j} = 0 \qquad \frac{1}{n-1} \sum_i (z_i^j - \overline{z^j})^2 = 1$$

- Какие объекты можно признать аномальными в базе Turkey Student Evaluation?
- Какую информацию можно извлечь из данных?
- Как можно использовать эту информацию в будущем?

Пример объектов - выбросов базы Turkey Evaluation Student

i	c	nb	att	diff	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	2	1	2	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3	3	2	2	1	1	1	1
1	2	1	3	4	5	5	4	4	5	5	4	4	5	5	5	4	5	5	4	4	5	5	5	4	4	5	5	4	4	4	5	4
1	2	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	1	1	1	3	2	2	2	2
1	2	1	2	4	5	3	3	3	2	2	3	3	3	4	4	5	5	4	3	3	3	4	2	2	4	4	5	5	4	4	5	5
1	2	1	1	2	1	1	1	1	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
1	2	1	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	3	3	3	3	3	3	3
1	2	1	2	3	1	1	1	1	2	2	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	2	1	3	4	3	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3	3	3	3	3	3	4	4	2	4	2	1	3
1	2	1	1	3	4	4	4	4	4	4	5	5	5	5	5	4	4	5	4	4	4	4	4	4	4	4	4	4	4	5	4	4
1	2	2	1	3	2	3	3	3	2	5	5	5	5	5	5	5	3	3	3	3	3	3	3	3	3	3	3	3	2	2	1	1	1
1	2	1	3	4	2	3	4	5	5	4	4	4	5	4	4	4	4	4	4	4	4	2	2	2	4	2	2	4	2	2	3	2
1	2	1	1	3	4	4	4	3	4	2	4	5	3	3	4	1	5	5	5	5	5	5	5	5	5	5	5	3	4	5	4	4	5
1	2	1	1	1	1	1	1	1	1	1	1	5	1	1	1	5	5	5	5	5	5	4	5	5	5	5	5	5	5	4	5	5
1	2	1	3	3	2	4	4	2	5	5	5	5	4	4	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5
1	2	1	4	3	3	5	4	4	5	5	4	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
1	2	1	3	3	4	4	3	3	3	3	3	3	3	3	3	5	4	4	3	3	4	3	3	3	3	3	3	3	3	3	3	3
1	2	1	4	3	4	4	4	3	4	3	4	4	4	4	4	4	5	5	5	5	5	5	4	4	4	4	4	4	4	4	4	4
1	2	1	0	1	3	3	1	3	1	2	2	2	2	1	1	1	3	4	4	3	2	4	1	3	3	3	2	3	4	2	3	3
1	2	1	1	5	5	2	2	2	5	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	5	4	5	5	1	5
1	2	1	1	4	3	3	3	4	4	4	4	3	3	4	4	4	4	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4
1	2	1	3	4	3	3	3	3	3	3	3	3	3	3	3	3	2	3	3	3	4	4	4	4	1	1	3	3	3	3	3	3

Поиск выбросов

- Ковариационная матрица близка к вырожденной (определитель ~0)
- Объекты в большинстве либо очень далеки от того чтобы быть выбросами, либо выбросы при практически любом уровне

Номер объекта	Расстояние Махалонобиса	Уровень отвержения
1	0.5697776	~0
3	1.5927358	~0
9	31.0974686	0.68
14	4.9708474	0.0000004
15	116.0313671	~1
113	0.4032532	~0
140	119.5756981	~1

Поиск выбросов

- Объекты-выбросы практически не меняются при разумном изменении параметра уровня значимости
- Объекты, которые были сочтены выбросами не выглядят аномальными
- В данном случае анализ многомерных выбросов не имеет смысла. Необходимо придумать критерий удаления

Практическое задание

1. Предложить методы анализа выбросов, учитывая особенности данных. Сделать анализ выбросов, удалить выбросы.
2. Проанализировать матрицу корреляций оценок по различным критериям качества преподавания. Выявить значимые корреляции. Объяснить высокие и низкие корреляции.
3. Сравнить матрицы корреляций для разных предметов.

Отслеживание времени работы программы

```
//Python
import time
t1 = time.time()
t2 = time.process_time()
…Algorithm execution…
print("Total time = ", time.time() - t1)
print("Processor time = ", time.process_time() - t2)



//R
t1 <- proc.time()
…Algorithm execution…
proc.time() - t1



//Java
long t1 = System.nanoTime();
…Algorithm execution…
System.out.println(System.nanoTime() - t1);
```