Практическая работа № 3. Методы машинного обучения для анализа данных

Работа № 3.1 Первичный анализ наборов данных

1. Источник данных: TurkeyStudentEvaluation

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

Свойства набора данных:

- 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)

Фрагмент источника данных

instr	class	nb.repeat	attendance	difficulty	Q1	Q2			Q19	Q20
1	2	1	3	5	3	3			3	3
1	2	1	3	4	3	3			3	3
1	2	1	0	1	5	5			5	5
1	2	1	3	5	3	3			3	3
1	2	1	3	4	5	5			5	5

Студенты оценивали параметры difficulty и Q1-Q28. Также приведена информация по посещаемости данным студентом некоторого занятия(attendance), какой раз данный студент проходит данный курс (nb.repeat), номер курса (class) и номер преподавателя (instr).

Описание атрибутов источника данных (на англ. из оригинала, на русском)

Attribute Information:

instr: Instructor's identifier; values taken from {1,2,3}
class: Course code (descriptor); values taken from {1-13}
repeat: Number of times the student is taking this course; values taken from {0,1,2,3,...}
attendance: Code of the level of attendance; values from {0, 1, 2, 3, 4}
difficulty: Level of difficulty of the course as perceived by the student; values taken from {1,2,3,4,5}
Q1: The semester course content, teaching method and evaluation system were provided at the start.
Q2: The course aims and objectives were clearly stated at the beginning of the period.
Q3: The course was worth the amount of credit assigned to it.
Q4: The course was taught according to the syllabus announced on the first day of class.
Q5: The class discussions, homework assignments, applications and studies were satisfactory.
Q6: The textbook and other courses resources were sufficient and up to date.
Q7: The course allowed field work, applications, laboratory, discussion and other studies.
Q8: The quizzes, assignments, projects and exams contributed to helping the learning.
Q9: I greatly enjoyed the class and was eager to actively participate during the lectures.
Q10: My initial expectations about the course were met at the end of the period or year.
Q11: The course was relevant and beneficial to my professional development.
Q12: The course helped me look at life and the world with a new perspective.
Q13: The Instructor's knowledge was relevant and up to date.
Q14: The Instructor came prepared for classes.

Q15: The Instructor taught in accordance with the announced lesson plan.
Q16: The Instructor was committed to the course and was understandable.
Q17: The Instructor arrived on time for classes.
Q18: The Instructor has a smooth and easy to follow delivery/speech.
Q19: The Instructor made effective use of class hours.
Q20: The Instructor explained the course and was eager to be helpful to students.
Q21: The Instructor demonstrated a positive approach to students.
Q22: The Instructor was open and respectful of the views of students about the course.
Q23: The Instructor encouraged participation in the course.
Q24: The Instructor gave relevant homework assignments/projects, and helped/guided students.
Q25: The Instructor responded to questions about the course inside and outside of the course.
Q26: The Instructor's evaluation system (midterm and final questions, projects, assignments, etc.) effectively measured the course objectives.
Q27: The Instructor provided solutions to exams and discussed them with students.
Q28: The Instructor treated all students in a right and objective manner.
Q1-Q28 are all Likert-type, meaning that the values are taken from {1,2,3,4,5}

Информация об полях источника данных:

instr: идентификатор инструктора; значения взяты из {1,2,3}
class: Код курса (дескриптор); значения взяты из {1-13}
repeat: сколько раз студент проходил этот курс; значения взяты из {0,1,2,3, ...}
attendance: Код уровня посещаемости; значения из {0, 1, 2, 3, 4}
difficulty: уровень сложности курса, который воспринимается студентом; значения взяты из {1,2,3,4,5}
Q1: Содержание семестрового курса, метод обучения и система оценивания были предоставлены в начале.
Q2: Цели и задачи курса были четко сформулированы в начале периода.
Q3: Курс стоил присвоенной ему суммы кредита.
Q4: Курс преподавался в соответствии с программой, объявленной в первый день занятий.
Q5: Обсуждения в классе, домашние задания, приложения и исследования были удовлетворительными.
Q6: Учебники и другие ресурсы курсов были достаточными и актуальными.
Q7: Курс допускал полевые работы, приложения, лабораторные, обсуждения и другие исследования.
Q8: Тесты, задания, проекты и экзамены способствовали обучению.
Q9: Мне очень понравился урок, и я очень хотел активно участвовать во время лекций.
Q10: Мои первоначальные ожидания относительно курса оправдались в конце периода или года.
Q11: Курс был актуален и полезен для моего профессионального развития.
Q12: Курс помог мне взглянуть на жизнь и мир с новой точки зрения.
Q13: Знания инструктора были актуальными и актуальными.
Q14: Инструктор прибыл подготовленным к занятиям.
Q15: Инструктор преподавал в соответствии с объявленным планом урока.
Q16: Инструктор был привержен курсу и был понятен.
Q17: Инструктор прибыл вовремя на занятия.
B18: Инструктор легко и четко произносит речь.
Q19: Инструктор эффективно использовал часы занятий.

Q20: Преподаватель объяснил курс и очень хотел помочь студентам.

Q21: Преподаватель продемонстрировал положительный подход к студентам.

Q22: Преподаватель был открыт и уважительно относился к мнению студентов о курсе.

Q23: Инструктор поощрял участие в курсе.

Q24: Преподаватель давал соответствующие домашние задания / проекты и помогал / руководил студентами.

B25: Инструктор ответил на вопросы о курсе внутри и вне курса.

Q26: Система оценки преподавателя (промежуточные и заключительные вопросы, проекты, задания и т. Д.) Эффективно измеряла цели курса.

Q27: Преподаватель предоставил решения к экзаменам и обсудил их со студентами.

Q28: Преподаватель относился ко всем студентам правильно и объективно.

Q1-Q28 все относятся к типу Лайкерта, что означает, что значения взяты из {1,2,3,4,5}

2. Описательные статистики

- Минимум и максимум
- Среднее значение
- Характеристики разброса
- Дисперсия
- Стандартное отклонение
- Интервал изменения
- Квантили
- Матрица ковариаций и корреляций (оценка связи между признаками)

Например: Вычисление описательных статистик NumPy:

```
import numpyas np
data = np.genfromtxt("turkiye.csv", delimiter=',')
for iin range(0,data.shape[1]):
# data.shape[1] размерность
    # второй мерности матрицы
    # (количество столбцов)
print("Признак ",i-2)
x = data[:,i]
    size = np.size(x)
    min = np.min(x)
    max = np.max(x)
print("Минимум: ",min)
print("Максимум: ",max)
    sum = np.sum(x)
print("Сумма: ",sum)
    sum2 = np.dot(x,x)
print("Суммаквадратов: ", sum2)
    mean = sum/size
    mean2 = sum2/size
print("Среднеезначение: ",mean)
print("Момент второго порядка: ", sum2/size)
var = np.var(x)
    SDM = var * size
print("Сумма квадратов отличий от средних: ", SDM)
print("Дисперсия: ",var)
    std = np.sqrt(var)
varcoef = std/mean
print("Среднеквадратичное отклонение: ",std)
print("Коэффициент вариации: ",varcoef)
print(" ")
corr = np.corrcoef(data, rowvar= 0)
print(np.max(corr))
```

```
print("Матрицакорреляции: ", corr)
covar = np.cov(data, rowvar= 0)
print("Матрицаковариаций: ", covar)
x = data[:,0]
print("Квантили: ",np.percentile(x,[25,50,75]))
```

Матрица корреляций признаков сложности предмета и всех вопросов.

	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

Рисунок 3.1. Матрица корреляций признаков сложности предмета и всех вопросов

Видно, что вопросы, расположенные рядом обладают как правило большей корреляцией, чем вопросы, расположенные в опроснике дальше друг от друга. Это говорит о том, что опросник составлялся так, чтобы вопросы были расположены по некоторому логическому порядку, возможно группами, которые раскрывают одну из сторон преподавания. Также видно, что сложность фактически не коррелирует с ответами на вопросы.

3. Аномалии в данных

Причины появления аномалий 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

Резюме: Аномалии в данных в разных наборах проявляются по-разному, выработать некоторый одинаковый подход сложно.

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

Из-за наличия пропусков в данных работа некоторых алгоритмов невозможна. Пути решения – удаление признаков и объектов с большим количеством пропусков, подстановка средних значений по признаку, ЕМ алгоритм подстановки.

В разных местах одни и те же «связующие» данные могут быть записаны в разной форме. Некоторые значения не могут оказаться правдой в силу разных причин (возраст вряд ли может составлять >150, возраст, рост, вес не могут быть отрицательными и для них известно в каких пределах эти данные могут логично располагаться). Некоторые значения могут выбиваться из общего ряда просто потому, что случилось некоторое редкое, маловероятное событие, которое было записано.

Разные числовые признаки могут располагаться в разных диапазонах.

Рассмотрим пример работы с пропущенными данными. Можно заметить, что очень много пропусков у признака 3, объектов 3, 13, 15.

Объект\признак	1	2	3	4	5	Число пропусков	% 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	
% пропусков	10	10	55	30	10		23

После удаления указанных признаков и объектов остаётся лишь некоторое число пропусков, которые можно попытаться восстановить.

Объект\признак	1	2	4	5	число пропусков	% 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

Объект\признак	1	2	4	5	число пропусков	% пропусков
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	
% пропусков	0	0	29.4	0		7.35

Пример аномальных данных в TurkeyStudentEvaluation.

Можно заметить большое количество объектов, где все ответы на вопросы одинаковые. Есть некоторые, где все ответы на вопросы кроме оценки сложности одинаковые. В опросниках часто бывает такая ситуация, что человеку лень или нет времени отвечать на вопросы, поэтому он просто ставит одинаковые галочки на все вопросы. Возникает вопрос, какие из объектов действительно являются лишними в смысле их недостоверности. Например, 2 человека не посещали занятия, указали сложность предмета как 1 (очень легкий) и ответили на все вопросы о преподавателе 1 (что в целом говорит об оценке преподавателя как очень плохо): как студенты, не посещавшие занятия, могли оценить преподавателя как плохого во всём? Каждый может придумать свой критерий, какие объекты являются лишними.

Таблица Пример аномальных данных в TurkeyStudentEvaluation

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

Методы поиска выбросов

- Поиск выбросов с использованием квартилей (данные выбросы ищутся по одному признаку):
  - 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 $[X_1, X_2]$:

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

- Поиск выбросов по распределениям признаков (данные выбросы могут быть найдены как многомерные – то есть целые объекты):
  - Все объекты, для которых выполнено неравенство, являются выбросами:

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

где Σ – матрица ковариаций признаков.

Стандартизация и нормализация данных

- Стандартизация задаёт чёткие границы, в которых располагаются значения некоторого признака.

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

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

  2) $z_i^j = \frac{x_i^j - a^j}{b^j - a^j}$, $z_i^j \in [0,1]$
- Нормализация задаёт свойства признака – мат. ожидание 0 и стандартное отклонение 1.

  Нормализация: $\mu^j = \overline{x^j}$, $\sigma_j^2 = \frac{1}{n-1}\sum_i \left(x_i^j - \overline{x^j}\right)^2$

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

Данные объекты были признаны выбросами (многомерными) среди первых 140 объектов (ответов студентов-посетителей курса №2). Можно заметить, что здесь присутствуют объекты с разными значениями признаков, причём существуют и более «крайние» объекты – из только единиц или пятерок – которые не попали в список выбросов. Применение метода поиска одномерных выбросов также не имеет смысла, поскольку зависит от значения квартилей. В некоторых случаях крайние оценки могут признаваться выбросом, если пользоваться этим методом, что не имеет смысла, поскольку если какая-то из оценок является выбросом, то нет никакого смысла использовать эту оценку при опросах.

Например, если k=1.5, $|Q_1-Q_3| = 2$, то можно показать, что все оценки будут признаны не выбросами. Если $|Q_1-Q_3| = 1$, то будет исключена как минимум 1 оценка из 5 (например, если $Q_1 = 2$, $Q_3 = 3$, то будет исключена оценка 5, если $Q_1 = 4$, $Q_3 = 5$, будут исключены 1 и 2). Если $|Q_1-Q_3| = 0$, будут исключены все оценки кроме $Q_1 = Q_3 =$ оценка.

Резюме:

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

Таблица. Пример объектов - выбросов базы TurkeyEvaluationStudent

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	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	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

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.
4. Проанализировать описательные статистики по преподавателям, разработать метод сравнения преподавателей по приведённым данным.
5. Проанализировать описательные статистики по предметам, разработать метод сравнения предметов по данным из набора.

Контрольные вопросы:

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