МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

«Дальневосточный федеральный университет»

(ДВФУ)

ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

(ШКОЛА)

Департамент информационных и компьютерных систем

ОТЧЕТ

по дисциплине «системы искусственного интеллекта»

Выполнил студенты группы Б9122-09.03.03пикд

Зверев Р. И.

Проверил преподаватель

Бочарова В. В.

зачтено/не зачтено

г. Владивосток

2025 г

Оглавление

Цель работы

Целью работы является первичный анализ данных и применение методов анализа для сравнительного анализа.

Постановка задачи

В данной работе рассматривается задача анализа данных на датасете Turkey Evaluation Student.

Необходимо реализовать следующие этапы и функции:

- Предложить методы анализа выбросов, учитывая особенности данных;
- Проанализировать матрицу корреляций оценок по различным критериям качества преподавания. Выявить значимые корреляции;
- Сравнить матрицы корреляций для разных предметов;
- Проанализировать описательные статистики по преподавателям, разработать метод сравнения преподавателей по приведённым данным;
- Проанализировать описательные статистики по предметам, разработать метод сравнения предметов по данным из набора.

Введение

В этой лабораторной работе выполняется первичный анализ набора данных Turkey Evaluation Student и сравнительный анализ полученных показателей. Задача включает разработку подходов к обнаружению и учёту выбросов, исследование корреляционных связей между оценками по критериям качества преподавания и сопоставление этих матриц для разных предметов. Кроме того, будут рассчитаны описательные статистики по преподавателям и по предметам, а также предложены методы их сравнения на основе имеющихся данных. Полученные результаты помогут выявить сильные и слабые стороны в оценках и понять, как предмет и преподаватель влияют на распределение оценок.

Описание данных

Описание набора данных GiveMeSomeCredit.

Для проведения анализа был выбран набор данных Turkiye Student Evaluation содержит ответы студентов на вопросы о качестве преподавания и содержит 5280 записей и 33 признака. Полное описание:

- **Instr**: идентификатор инструктора, значения взяты из {1,2,3};
- **Class**: код курса (дескриптор), значения взяты из {1-13};
- **Nb.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;
- **Q18**: 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
- **Q25**: 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 относятся к типу Лайкерта, что означает, что значения взяты из {1,2,3,4,5}.

Анализ выбросов

В качестве способа удаления аномалий в датасете я предлагаю рассмотреть 2 способа: Isolation Forest и Local Outliner Factor (LOF).

Обычные алгоритмы машинного обучения, например, SVM или нейросети, пытаются описать нормальное распределение данных, а затем искать выбросы. Isolation Forest идёт с другого конца: он не строит плотностную модель, а просто пытается изолировать выбросы.

Как это происходит:

- Строим дерево, где каждый узел случайно выбирает один признак и случайное значение разбиения;
- Рекурсивно делим данные, пока каждая точка не окажется в своём отдельном листе;
- Считаем аномальность точки по тому, насколько быстро она была изолирована (чем короче путь, тем аномальнее).

Коэффициент локального выброса (LOF) - это основанный на плотности неконтролируемый алгоритм машинного обучения, используемый для выявления аномальных точек данных путем сравнения локальной плотности точки с локальными плотностями ее соседей.

Принцип работы:

- **Оценка локальной плотности:** LOF вычисляет локальную плотность вокруг каждой точки данных, сравнивая плотность соседних точек. Для этого используется конкретная точка и определенное окружение вокруг нее;
- **Сравнение плотностей:** алгоритм вычисляет коэффициент локального выброса, сравнивая плотность интересующей точки с плотностями ее соседей. Если плотность точки намного ниже, чем у соседних, то, скорее всего, это выброс;
- **Оценка и пороговое значение:** LOF присваивает оценку каждой точке данных, указывая степень ее “отклонения”. Эта оценка может быть установлена как пороговая, чтобы классифицировать точки как нормальные или аномальные.

В итоге Isolation Forest с параметрами n_estimators = 1000, max_samples = 1000 нашел 1200 записей как аномальные, а LOF с параметрами n_neighbors = 150, leaf_size = 50 нашел 932 записи как аномальные.

Объединяем результаты двух алгоритмов и удаляем из исходного датасета аномалии. В итоге из 5820 записей остается 5091 запись.

Для оценки преподавателей и курса можно сделать предположение, что человек, который впервые проходит курс и ни разу не посещал занятия (nb.repeat = 1 и attendance = 0) не может дать адекватную оценку. Такие записи удаляем. Таких записей 1468 штук. После удаления из датасета остается 3623 записи.

Сравнение корреляций

Матрица корреляций признаков всего датасета выглядит так:

Как можно заметить, сильные корреляции у кол-ва прохождений курса (nb.repeat) и посещаемостью (attendance). Ответы на вопросы заметно коррелируют с количеством прохождений курса и посещаемостью. Сложность курса никак не влияет на вопросы. Красная область показывает, что вопросы связаны логическим порядком.

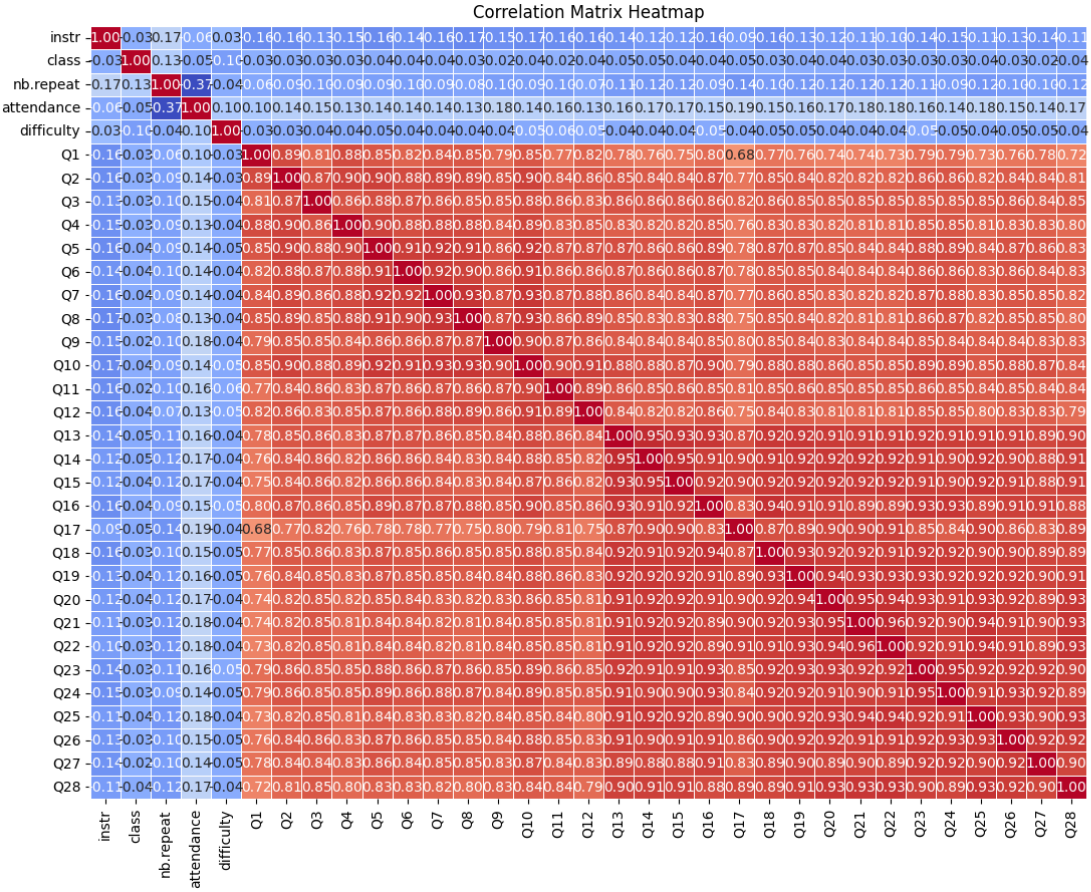

Рисунок 1. Матрица корреляций датасета

Сравнение матрицы корреляций разных предметов

Рассмотрим матрицу корреляций предмета №1:

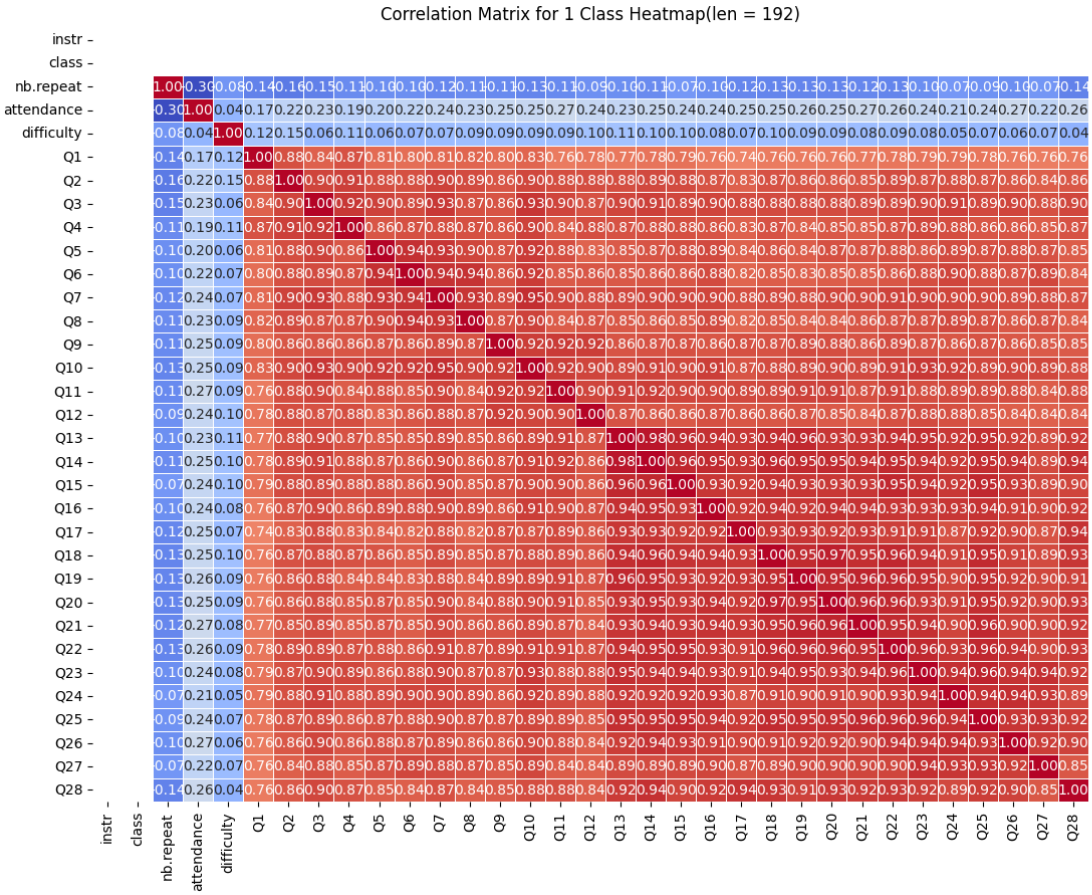

Рисунок 2. Матрица корреляций для предмета №1

Проверим еще матрицу для предмета №8, например:

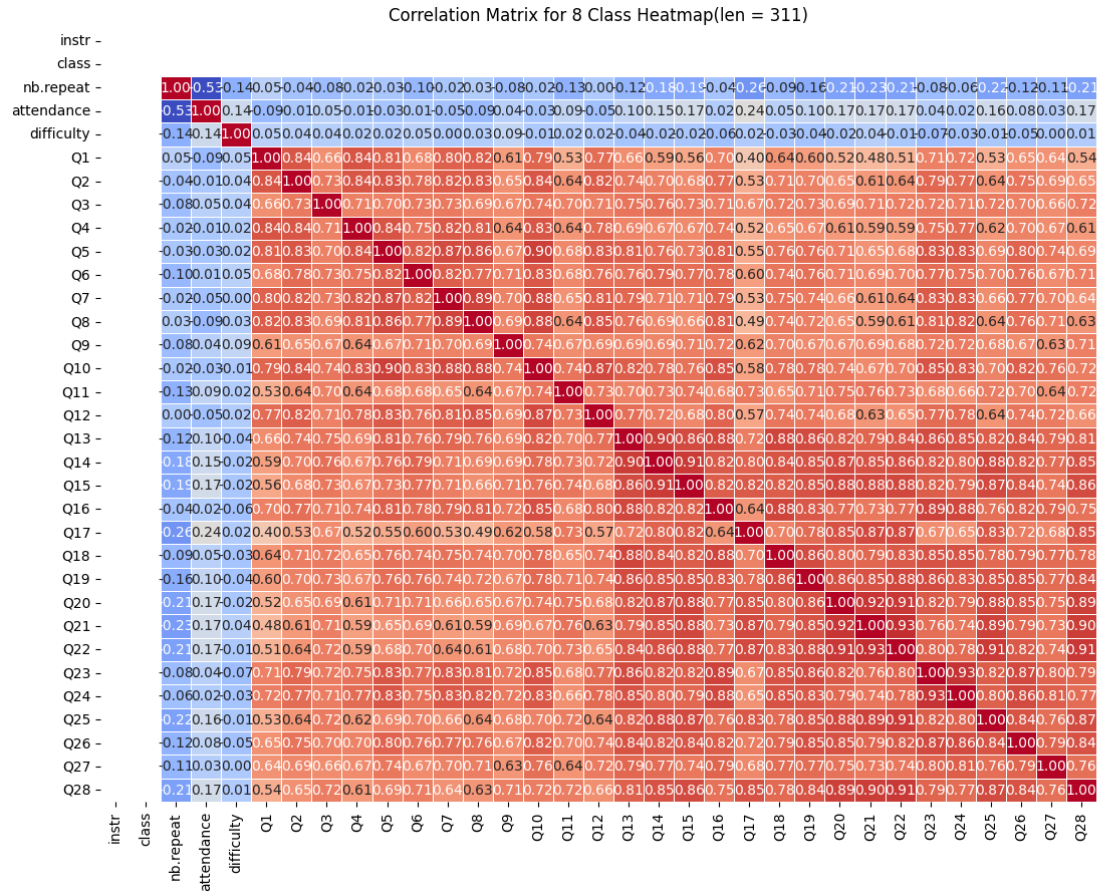

Рисунок 3. Матрица корреляций для предмета №8

Матрица предмета №13:

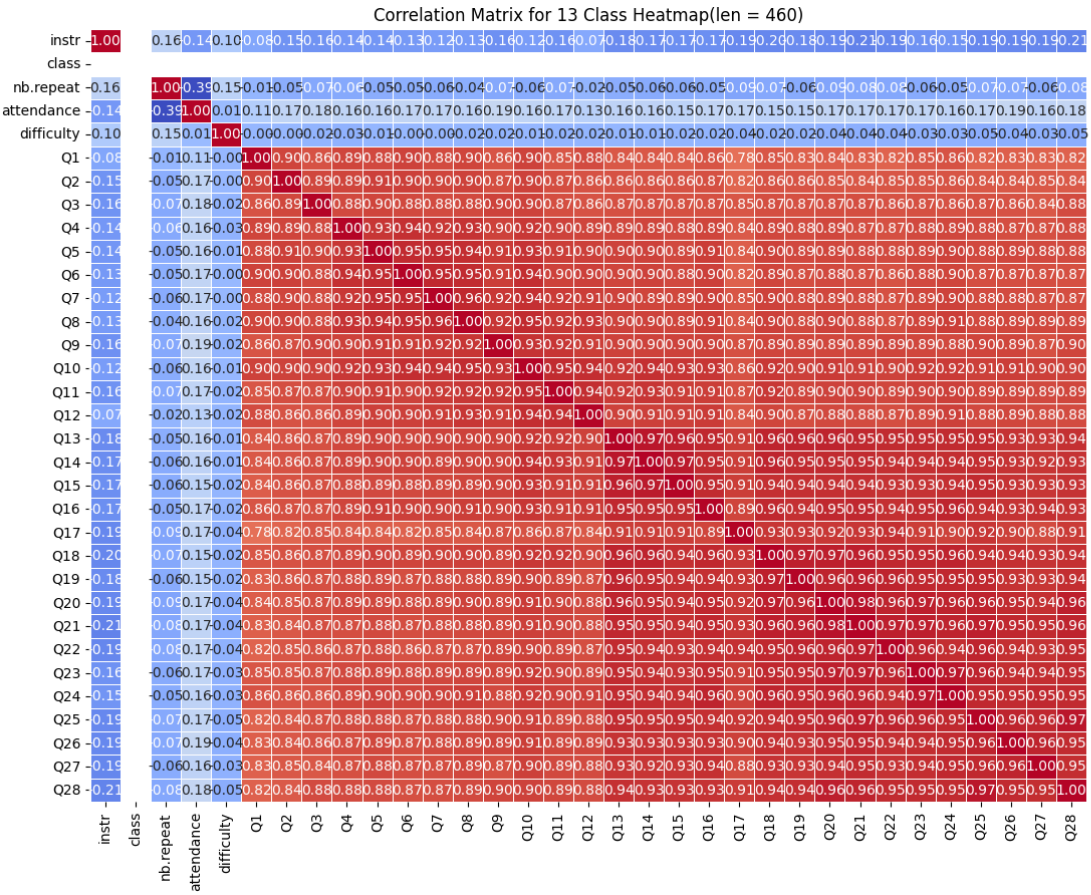

Рисунок 4. Матрица корреляций предмета №13

Можно заметить, что у всех предметов есть корреляция между количеством прохождений курса и посещением курса. Также, на ответы на вопросы влияет посещение курса. У некоторых предметов можно выделить слабо значимые вопросы, например у предмета №8 вопрос 17. Еще можно заметить, что на ответы на некоторые вопросы влияет количество прохождений курса.

Метод сравнения преподавателей

Для сравнения преподавателей будет считать аггрегированную оценку по вопросам 13 – 28, так как они связаны с мнением о преподавателях.

Такая статистика показывает следующий результат:

Таблица 1. Аггрегированная статистика преподавателей

Преподаватель	**Общее число ответов**	**Среднее значение**	**Медиана**	**Доля хороших оценок (4, 5)**
1	544	3.559	4	0.5404
2	952	3.575	4	0.5147
3	2127	3.226	3	0.3521

Также, построим для наглядности распределения оценок «ящик с усами»:

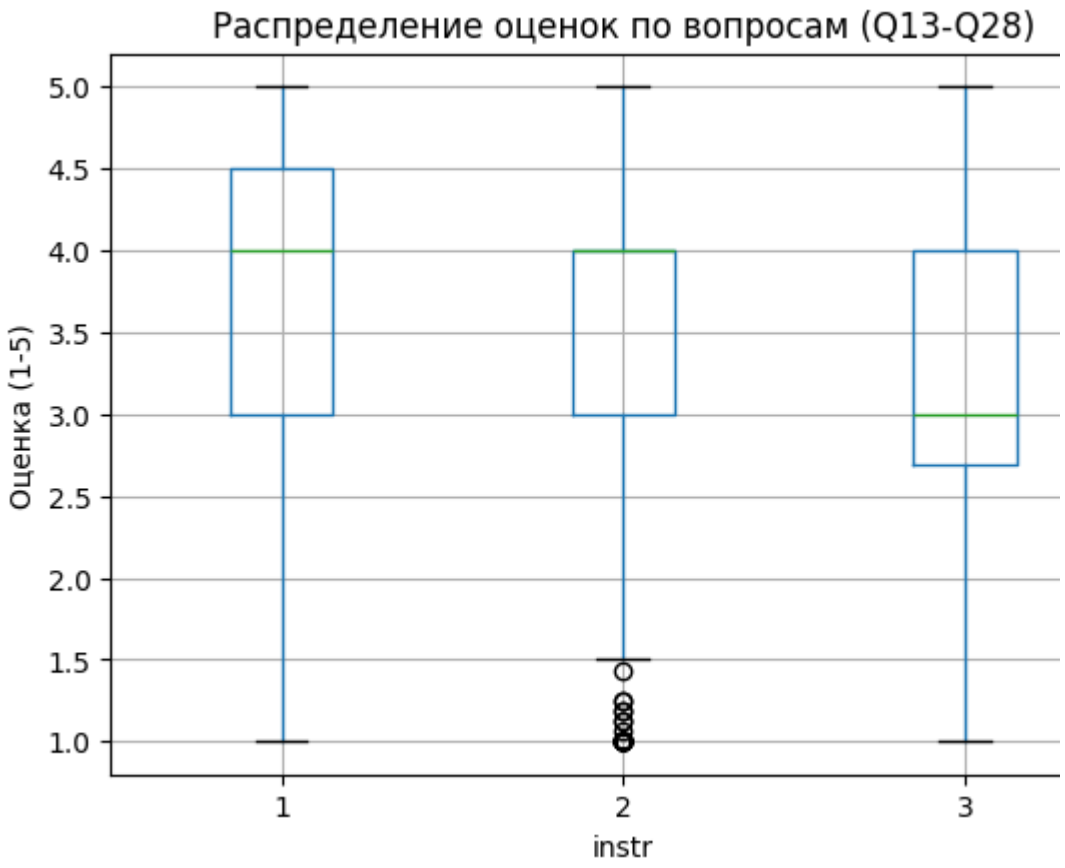

Рисунок 5. Box Plot преподавателей

По таким данным видно, что у преподавателя 1 и 2 хорошая доля оценок 4 и 5 при куда меньшем количестве записей, чем у преподавателя 3. Но у преподавателя 2 оценки немного не стабильны, чем у остальных.

Сравнение предметов

Проведем похожую статистику для предметов, смотря на все вопросы, так как мнение о предмете, так же, складывается от преподавателя.

Аггрегированная статистика выглядит так:

Таблица 2. Аггрегированная статистика предметов

Предмет	Кол-во записей	Среднее значение	Медиана	Доля хороших оценок
1	192	3.600	3.9285	0.479
2	105	3.785	4	0.600
3	535	3.088	3	0.290
4	92	2.919	3	0.120
5	396	3.356	3.3571	0.328
6	389	3.470	3.75	0.416
7	132	3.181	3.1607	0.311
8	311	3.313	3.3571	0.299
9	342	3.194	3	0.263
10	307	3.607	4	0.541
11	329	3.478	3.5714	0.410
12	33	3.182	3.178571	0.333
13	460	3.007	3	0.309

Рассмотрим распределение оценок:

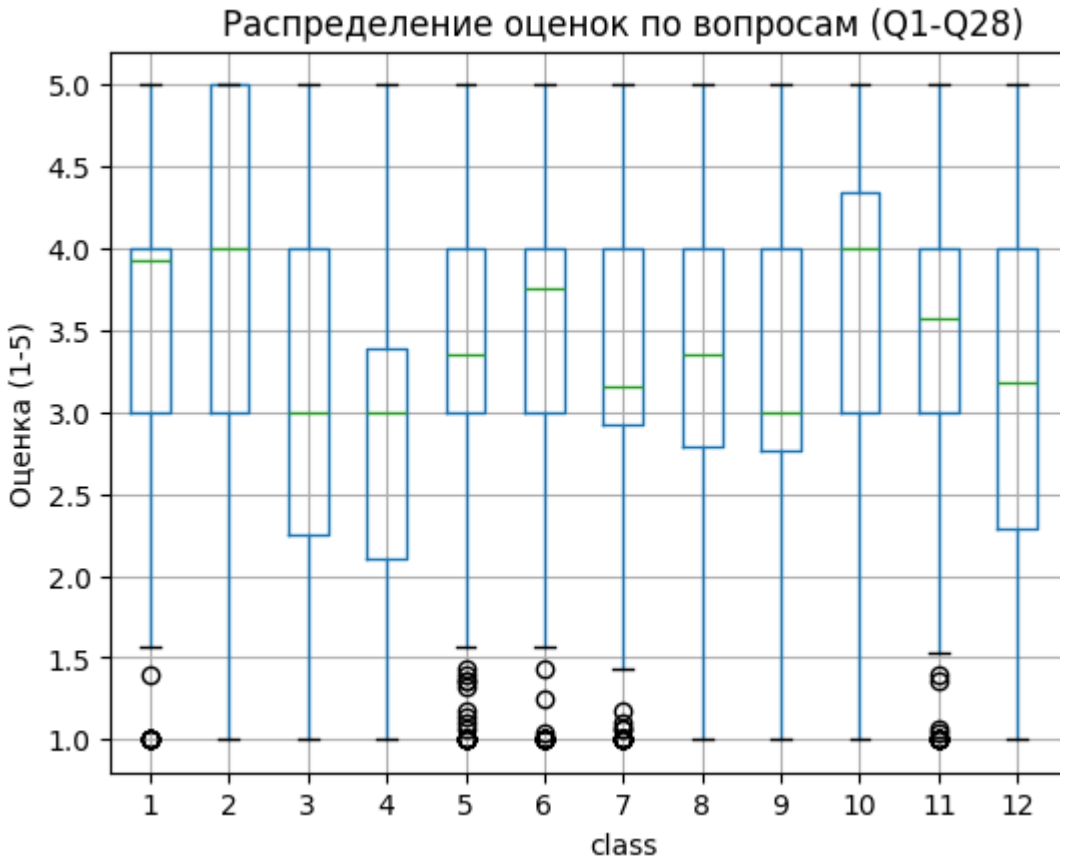

Рисунок 6. Распределение оценок по предметам

Как видно, ученики предпочитают предметы 2 и 10 всем остальным, на топ-3 стоит предмет №1. Предметы 4 и 12 сравнивать не стоит, так как у них очень мало записей, чем у остальных. Самыми не любимыми являются предметы 3, 9 и 13.

Заключение

В работе реализованы два метода обнаружения аномалий — Isolation Forest (n_estimators=1000, max_samples=1000), выявивший 1200 аномалий, и LOF (n_neighbors=150, leaf_size=50), выявивший 932 аномалии; их объединение и удаление даёт 5091 запись из исходных 5820. Дополнительная фильтрация записей с nb.repeat = 1 и attendance = 0 (1468 записей) оставляет в итоговом датасете 3623 наблюдения.

В матрице корреляций всего датасета наблюдаются сильные связи между количеством прохождений курса (nb.repeat) и посещаемостью (attendance); ответы на вопросы заметно коррелируют с этими переменными, в то время как сложность курса не проявляет значимой связи с вопросами.

Сравнение корреляционных матриц по предметам показывает сходную закономерность: для большинства предметов присутствует корреляция между nb.repeat и attendance, а посещаемость влияет на ответы по вопросам; у отдельных предметов видны слабые специфические взаимосвязи (например, вопрос 17 у предмета №8). Аггрегированная статистика по преподавателям (вопросы 13–28) даёт следующие результаты: преподаватель 1 — среднее 3.559, доля оценок 4–5 = 0.5404 (544 ответов); преподаватель 2 — среднее 3.575, доля 0.5147 (952 ответа); преподаватель 3 — среднее 3.226, доля 0.3521 (2127 ответов). Аггрегированная статистика по предметам показывает наилучшие средние и доли «хороших» оценок у предметов 2 и 10, тогда как предметы 3, 9 и 13 имеют наиболее низкие показатели; предметы 4 и 12 имеют малые объёмы выборки.

В целом данные демонстрируют устойчивые связи между повторностью прохождения, посещаемостью и оценками, а также различия в восприятии преподавателей и предметов.

Список литературы

1. GitHub: исходный код лабораторной работы. – URL: Лабораторная работа №3.1 (дата обращения: [10.10.2025]). – Текст: электронный.