МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Московский государственный технический университет имени Н.Э. Баумана

(национальный исследовательский университет)»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**по курсу**

«Data Science Pro»

Слушатель Строкач Екатерина Андреевна

Москва, 2025

Оглавление

[Введение 3](#_Toc199528783)

[**1 Аналитическая часть** 4](#_Toc199528784)

[1.1 Постановка задачи 4](#_Toc199528785)

[1.2 Описание используемых методов 15](#_Toc199528786)

[1.2.1 Ансамбль из множества деревьев решений 16](#_Toc199528787)

[1.2.2 Метод K-ближайших соседей (K-Nearest Neighbors, KNN) 17](#_Toc199528788)

[1.2.3 Метод Gradient Boosting 18](#_Toc199528789)

[1.2.4 Метод с поддерживающими векторами (Support Vector Regression, SVR) 18](#_Toc199528790)

[1.3 Разведочный анализ данных 19](#_Toc199528791)

[1.3.1 Выбор признаков 20](#_Toc199528792)

[1.3.2 Препроцессинг 20](#_Toc199528793)

[1.3.3 Поиск гиперпараметров по сетке 21](#_Toc199528794)

[1.3.4 Метрики качества моделей 21](#_Toc199528795)

[2 Практическая часть 22](#_Toc199528796)

[2.1. Разбиение и предобработка данных 22](#_Toc199528797)

[2.1.1 Для прогнозирования модуля упругости при растяжении 22](#_Toc199528798)

[2.1.2 Для прогнозирования прочности при растяжении 22](#_Toc199528799)

[Заключение 23](#_Toc199528800)

# Введение

Целью данной работы является создание методов прогнозирования характеристик новых композитов. Композитные материалы представляют собой сложные системы, состоящие из нескольких компонентов, которые четко разграничены между собой, сохраняя свои индивидуальные свойства. Такие материалы обладают неоднородной структурой, что позволяет достигать уникальных эксплуатационных характеристик, недостижимых для каждого компонента по отдельности.

Основу композитов составляет матрица — базовый элемент, в который добавляются армирующие компоненты, называемые наполнителями. Наполнитель распределяется внутри матрицы равномерно или в определённой ориентации, образуя сложную пространственную структуру. Эта структура определяет механические и физические свойства материала.

Ключевая особенность композитов заключается в их адаптируемости. Изменяя долю наполнителя, его геометрию, пространственное распределение и характеристики межфазной границы, можно получать материалы с заданными свойствами. Это делает их востребованными в высокотехнологичных отраслях:

- авиация и космическая техника;

- металлургия и добывающая промышленность;

- автомобилестроение и машиностроение;

- электроника и ядерная энергетика;

- медицина и строительство.

Высокая стоимость разработки композитов обусловлена сложностью их состава и необходимости многочисленных испытаний для получения требуемых характеристик. Поскольку предсказать свойства композитов на основе исходных компонентов затруднительно, производителям приходится проводить экспериментальные исследования, что увеличивает затраты времени и ресурсов.

Использование методов машинного обучения позволяет ускорить процесс создания новых материалов. Системы прогнозирования, основанные на анализе данных, могут минимизировать количество физических экспериментов, одновременно повышая точность и скорость проектирования. Это делает предложенный подход крайне актуальным в условиях растущего спроса на композитные материалы с уникальными свойствами.

# **1 Аналитическая часть**

# 1.1 Постановка задачи

Целью данной работы является исследование композитного материала, состоящего из базальтопластиковой матрицы и углепластиковых нашивок. Основной задачей является разработка моделей машинного обучения для прогнозирования конечных свойств данного материала на основе входных данных. Кроме того, предполагается создание удобного приложения, которое позволит использовать построенные модели специалистам в предметной области.

Исходные данные: для решения задачи был предоставлен датасет, содержащий информацию о свойствах матрицы и наполнителя, производственных параметрах и характеристиках готового композита. Данные представлены в виде двух отдельных файлов:

1. Файл X\_bp (свойства базальтопластика):

- Количество признаков: 10 (плюс индекс);

- Объем данных: 1023 строки.

2. Файл X\_nup (свойства углепластика):

- Количество признаков: 3 (плюс индекс);

- Объем данных: 1040 строк.

Файлы объединены с использованием операции INNER по методу JOIN. В результате объединения часть строк из файла X\_nup была исключена, для них отсутствовали соответствующие данные в файле X\_bp. Итоговый датасет содержит:

- Количество признаков: 13 (плюс индекс);

- Объем данных: 1023 строки.

Характеристики датасета:

- Тип данных: Все признаки, кроме одного, представлены в формате float64 (вещественные числа), и один признак в формате int64.

- Пропуски: Отсутствуют.

Типы признаков:

- Все признаки, за исключением "Угол нашивки", являются количественными и принимают непрерывные значения.

- Признак "Угол нашивки" принимает два значения и является категориальным.

Объединенный датасет представляет собой, в целом, чистый набор данных, не требующий предварительной обработки, связанной с преобразованием типов или устранением пропусков. Эти данные станут основой для построения прогнозных моделей. Характеристики датасета указана в таблице 1.

|  |  |  |  |
| --- | --- | --- | --- |
| № | Параметр | Кол-во строк | Тип данных |
| 1 | Угол нашивки, град | 1023 | int64 |
| 2 | Шаг нашивки | 1023 | float64 |
| 3 | Плотность нашивки | 1023 | float64 |
| 4 | Соотношение матрица-наполнитель | 1023 | float64 |
| 5 | Плотность, кг/м3 | 1023 | float64 |
| 6 | модуль упругости, ГПа | 1023 | float64 |
| 7 | Количество отвердителя, м.% | 1023 | float64 |
| 8 | Содержание эпоксидных групп,%\_2 | 1023 | float64 |
| 9 | Температура вспышки, С\_2 | 1023 | float64 |
| 10 | Поверхностная плотность, г/м2 | 1023 | float64 |
| 11 | Модуль упругости при растяжении, ГПа | 1023 | float64 |
| 12 | Прочность при растяжении, МПа | 1023 | float64 |
| 13 | Потребление смолы, г/м2 | 1023 | float64 |

На основе статистических данных, которые приведены на рисунке 1, по столбцам таблицы можно сделать следующие выводы:

1. Соотношение матрица-наполнитель:

- Среднее значение составляет 2.93, с отклонением (стандартное отклонение) 0.91.

- Минимальное значение 0.39, максимальное — 5.59, что указывает на широкое распределение значений.

- Большинство значений сосредоточено между 2.32 и 3.55.

2. Плотность, кг/м³:

- Среднее значение плотности — 1975.73 кг/м³, с небольшим стандартным отклонением (73.73).

- Плотность варьируется от 1731.76 до 2207.77 кг/м³, с 50% значений, лежащих в диапазоне от 1924.16 до 1977.62 кг/м³.

3. Модуль упругости, ГПа:

- Среднее значение модуля упругости — 739.92 ГПа с очень высоким стандартным отклонением (330.23 ГПа).

- Диапазон значений от 2.44 до 1911.54 ГПа, что свидетельствует о большом разнообразии в материалах или измерениях.

- 50% значений сосредоточено между 500.05 и 961.81 ГПа.

- Количество отвердителя, м.%:

- Среднее количество отвердителя — 110.57%, с отклонением 28.30%.

- Диапазон значений от 17.74% до 198.95%, что может свидетельствовать о различной концентрации отвердителя в материалах.

4. Содержание эпоксидных групп, %:

- Среднее значение — 22.24%, с отклонением 2.41%.

- Значения варьируются от 14.25% до 33%, что говорит о разнообразии химического состава.

5. Температура вспышки, °C:

- Среднее значение температуры вспышки — 285.88°C, с отклонением 40.94°C.

- Диапазон температур от 100°C до 413.27°C, что свидетельствует о возможном широком спектре химических веществ или различных условиях тестирования.

6. Поверхностная плотность, г/м²:

- Среднее значение поверхностной плотности — 482.73 г/м², но есть значительное отклонение (281.31 г/м²).

- Большинство значений сосредоточено в диапазоне от 266.82 до 693.23 г/м².

7. Модуль упругости при растяжении, ГПа:

- Среднее значение — 73.33 ГПа, с отклонением 3.12 ГПа.

- Диапазон значений от 64.05 до 82.68 ГПа, что говорит о сравнительно стабильных материалах в этом параметре.

8. Прочность при растяжении, МПа:

- Среднее значение прочности — 2466.92 МПа, с отклонением 485.63 МПа.

- Прочность варьируется от 1036.86 МПа до 3848.44 МПа, что указывает на значительные различия в прочности материалов.

9. Потребление смолы, г/м²:

- Среднее значение потребления смолы — 218.42 г/м², с отклонением 59.74 г/м².

- Потребление смолы варьируется от 33.80 г/м² до 414.59 г/м².

10. Угол нашивки, град:

- Среднее значение — 44.25°, с высоким стандартным отклонением (45.02°).

- Углы нашивки варьируются от 0° до 90°, при этом 50% значений имеют угол 0°.

11. Шаг нашивки:

- Среднее значение шага нашивки — 6.90, с отклонением 2.56.

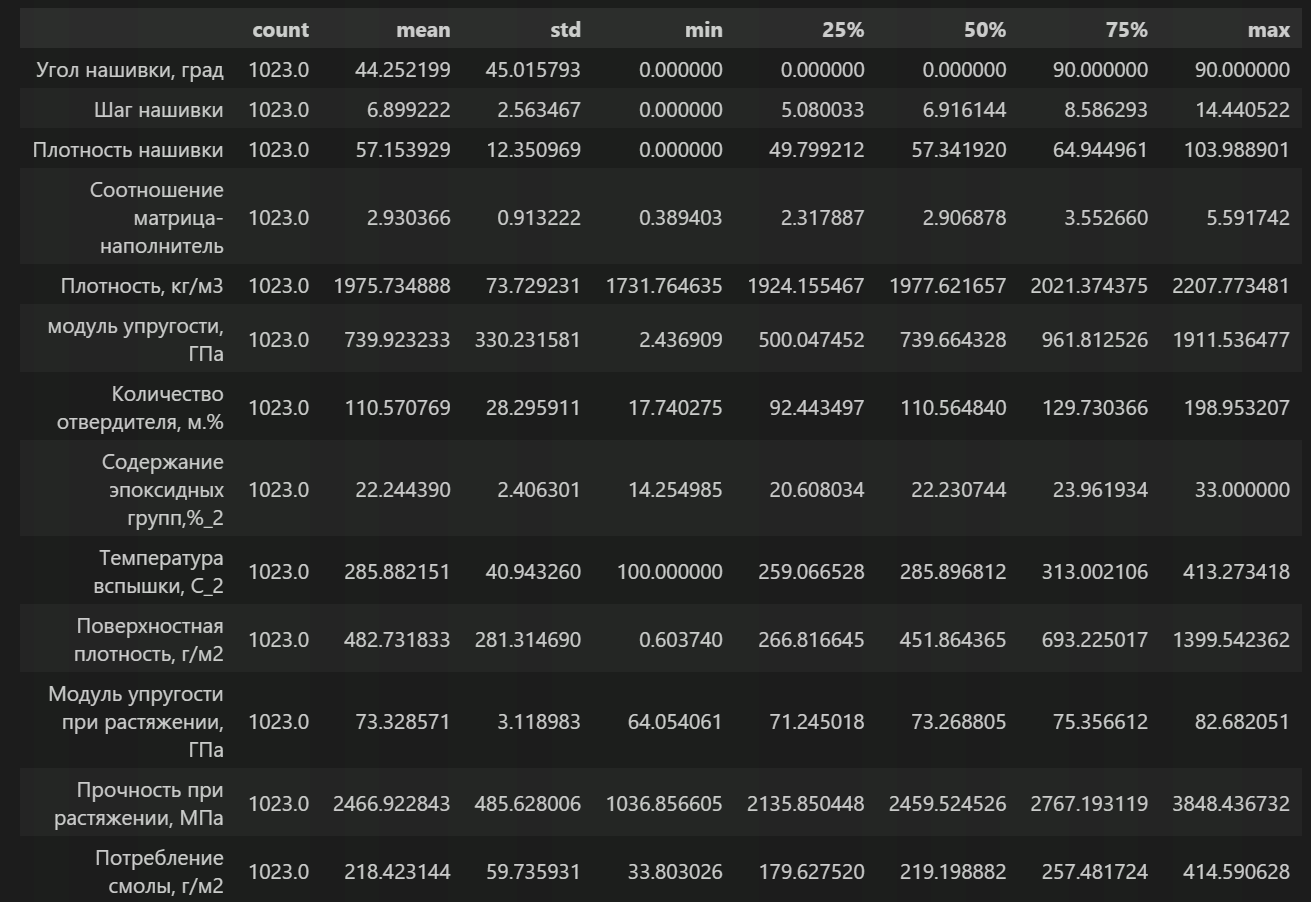
- Шаг нашивки варьируется от 0 до 14.44, что может свидетельствовать о разных типах тканей или изделий.

12. Плотность нашивки:

- Среднее значение плотности нашивки — 57.15 г/м², с отклонением 12.35.

- Плотность варьируется от 0 до 103.99 г/м², что также указывает на разнообразие в характеристиках материалов.

Рисунок 1



Общие выводы:

Данные имеют разброс, в особенности для таких параметров, как модуль упругости, прочность при растяжении и плотность нашивки.

Для большинства параметров характерна нормальная или близкая к нормальной дисперсия, что может свидетельствовать о стабильности данных с небольшими вариациями.

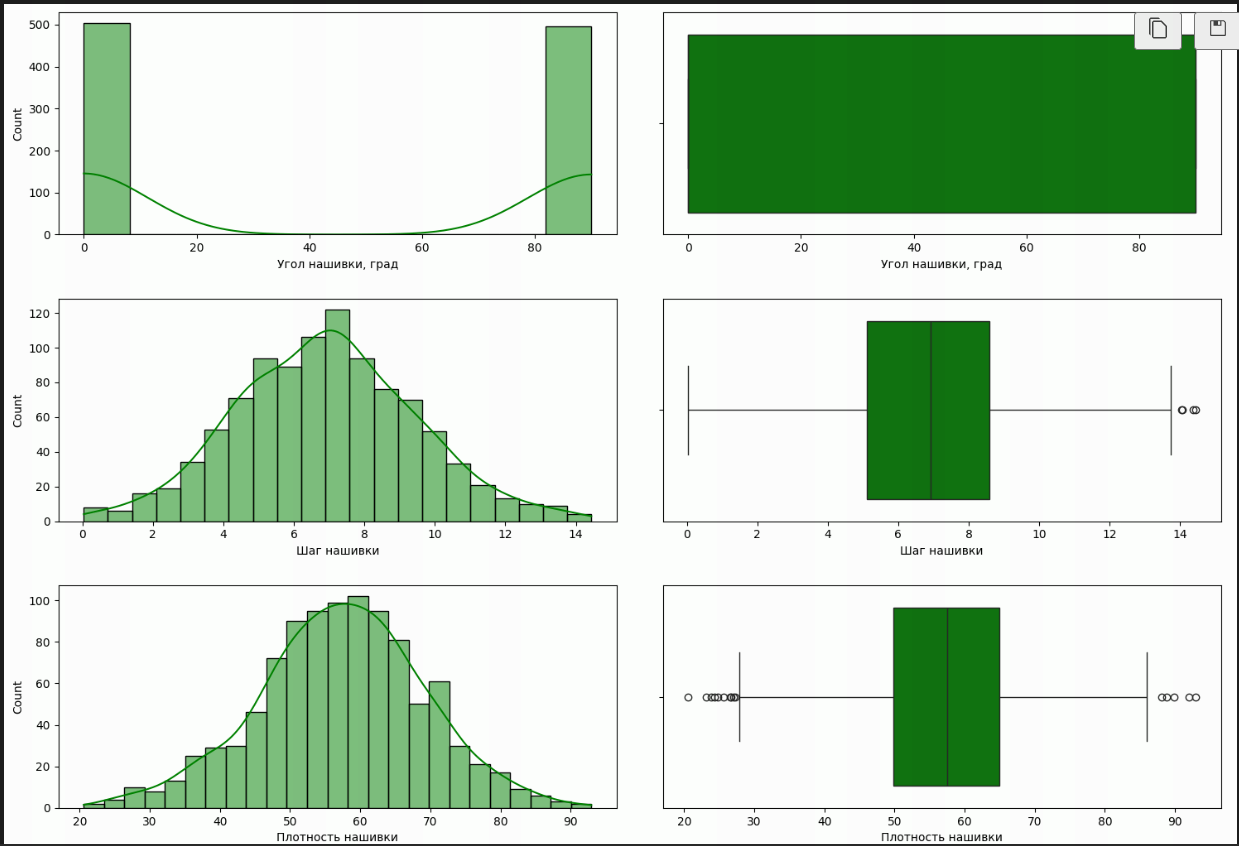
Некоторые параметры (например, "Угол нашивки") могут требовать дополнительной проверки на наличие аномальных значений (например, нулевые углы или слишком большие значения).

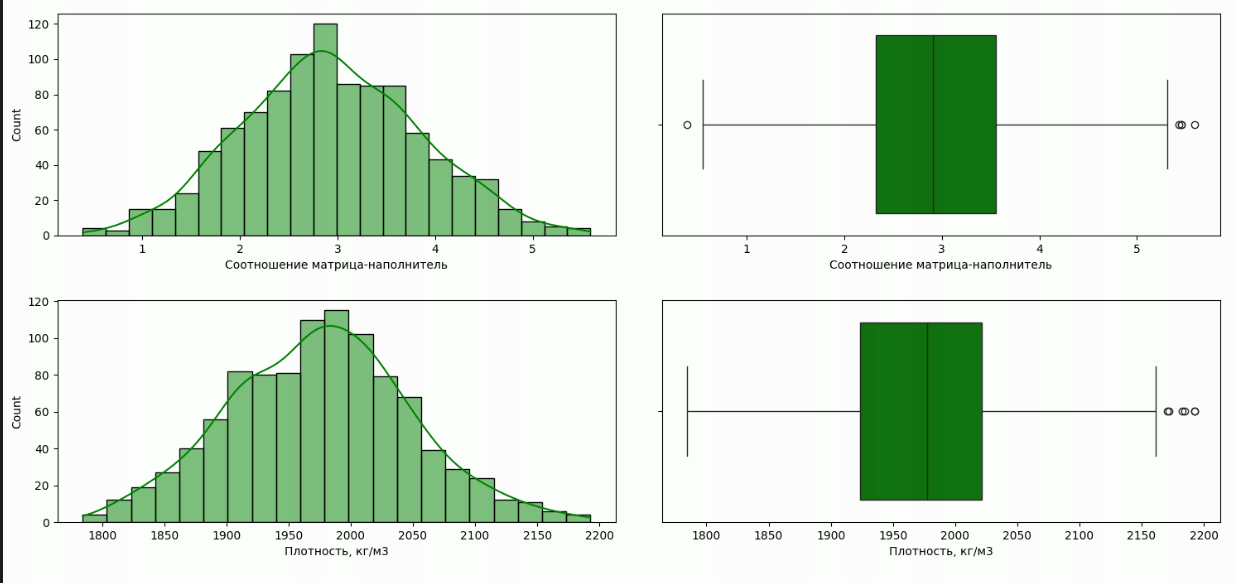
Некоторые параметры демонстрируют широкий диапазон значений, такие как модуль упругости и поверхностная плотность, что может свидетельствовать о гетерогенности материалов. В то же время в данных могут присутствовать выбросы, особенно в параметрах, связанных с температурой вспышки, модулем упругости и поверхностной плотностью.

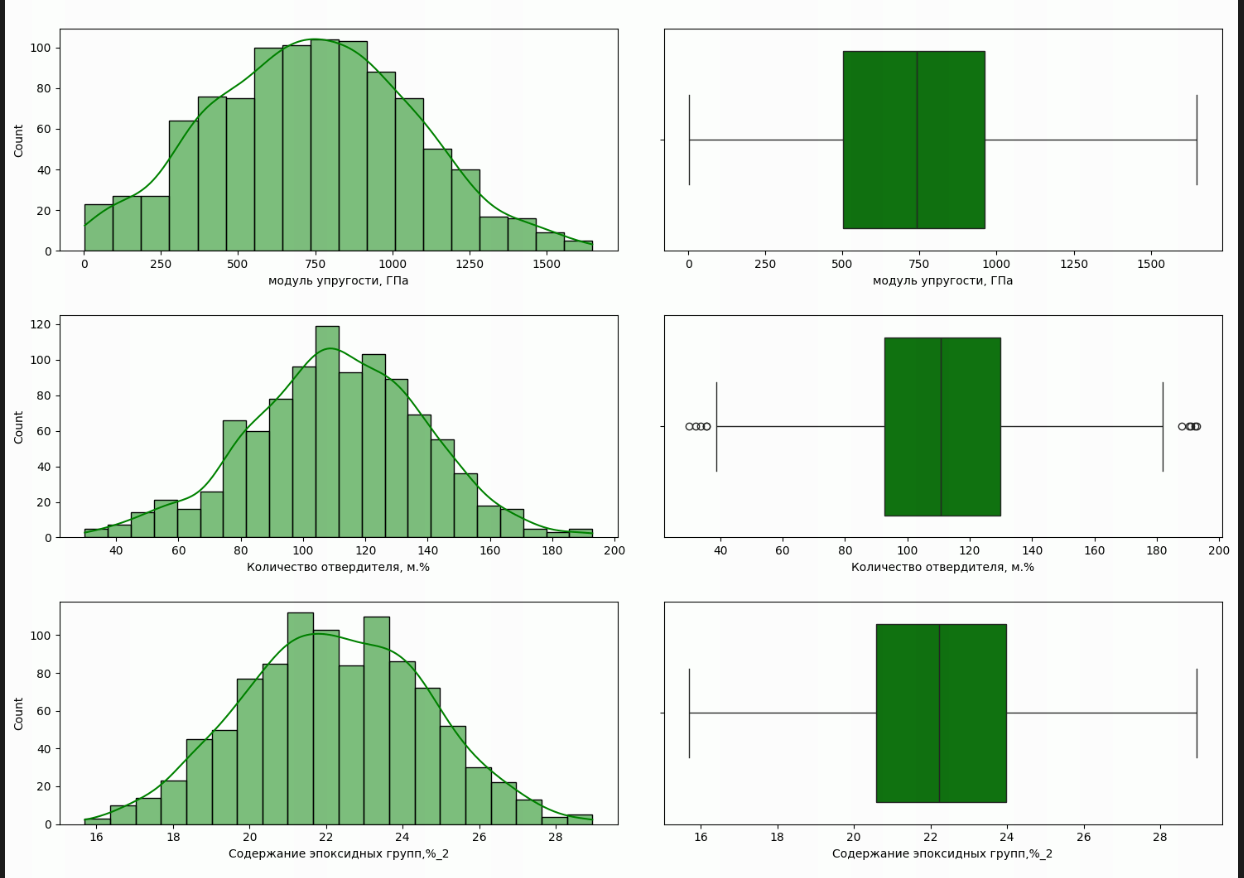
Параметры, такие как модуль упругости при растяжении и прочность при растяжении, имеют относительно узкие диапазоны значений и, вероятно, могут быть хорошо предсказаны с использованием методов машинного обучения.

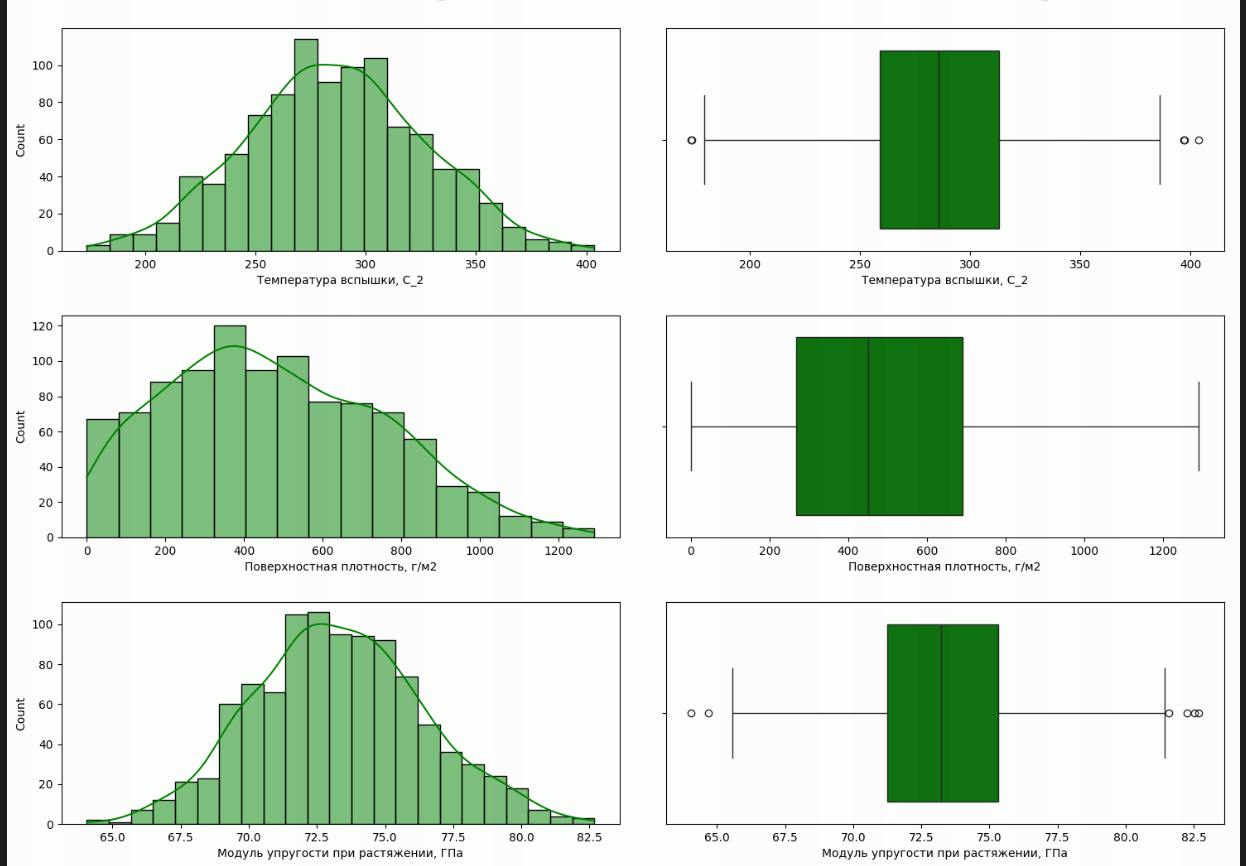
Для более детального анализа распределений и выявления возможных выбросов выполнена визуализация данных, такая как построение гистограмм и диаграмм типа «ящик с усами». Эти графики, представленные на рисунках 2–4, показали, что все признаки, кроме «Угол нашивки», имеют распределение, близкое к нормальному, и принимают только неотрицательные значения. Признак «Угол нашивки» является категориальным и принимает строго два значения: 0 и 90.

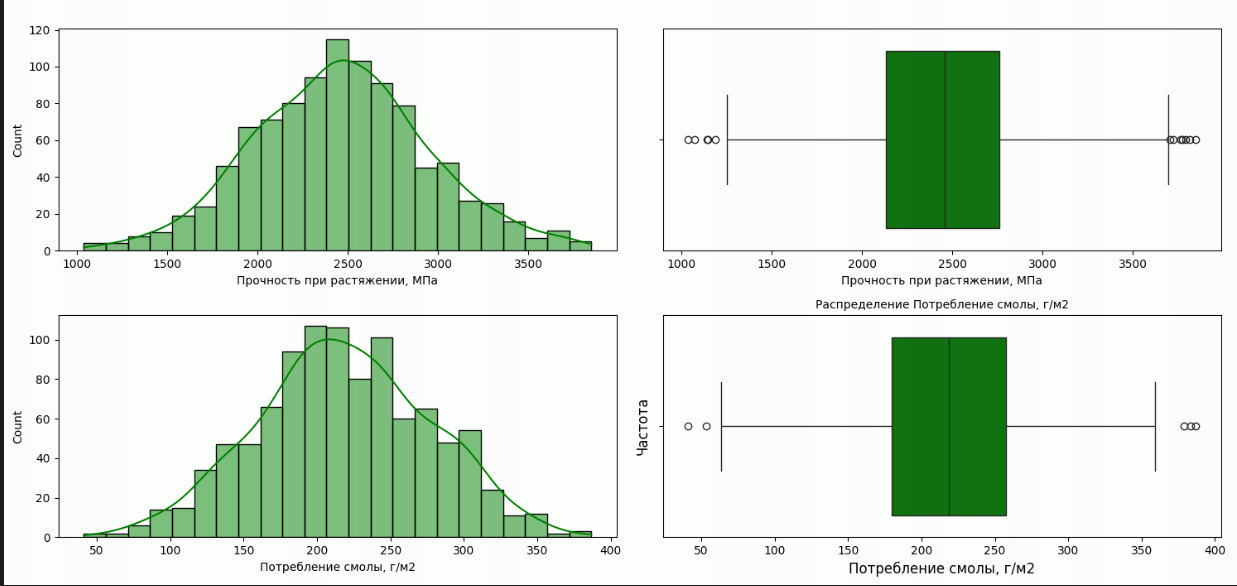
Рисунок 2 – Графики распределения всех типов столбцов

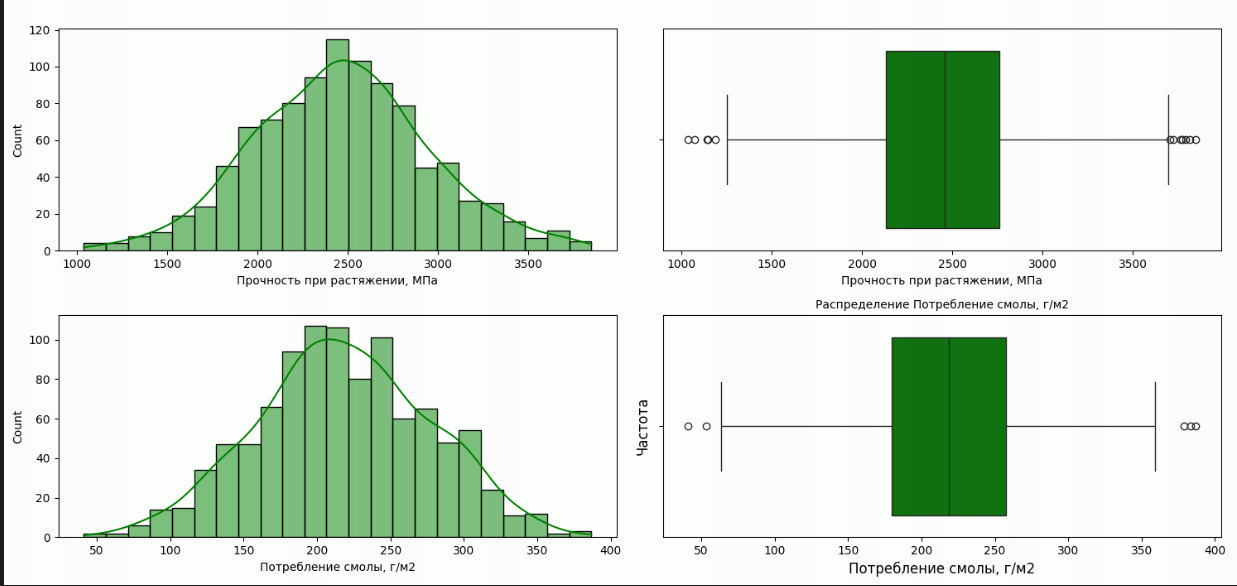








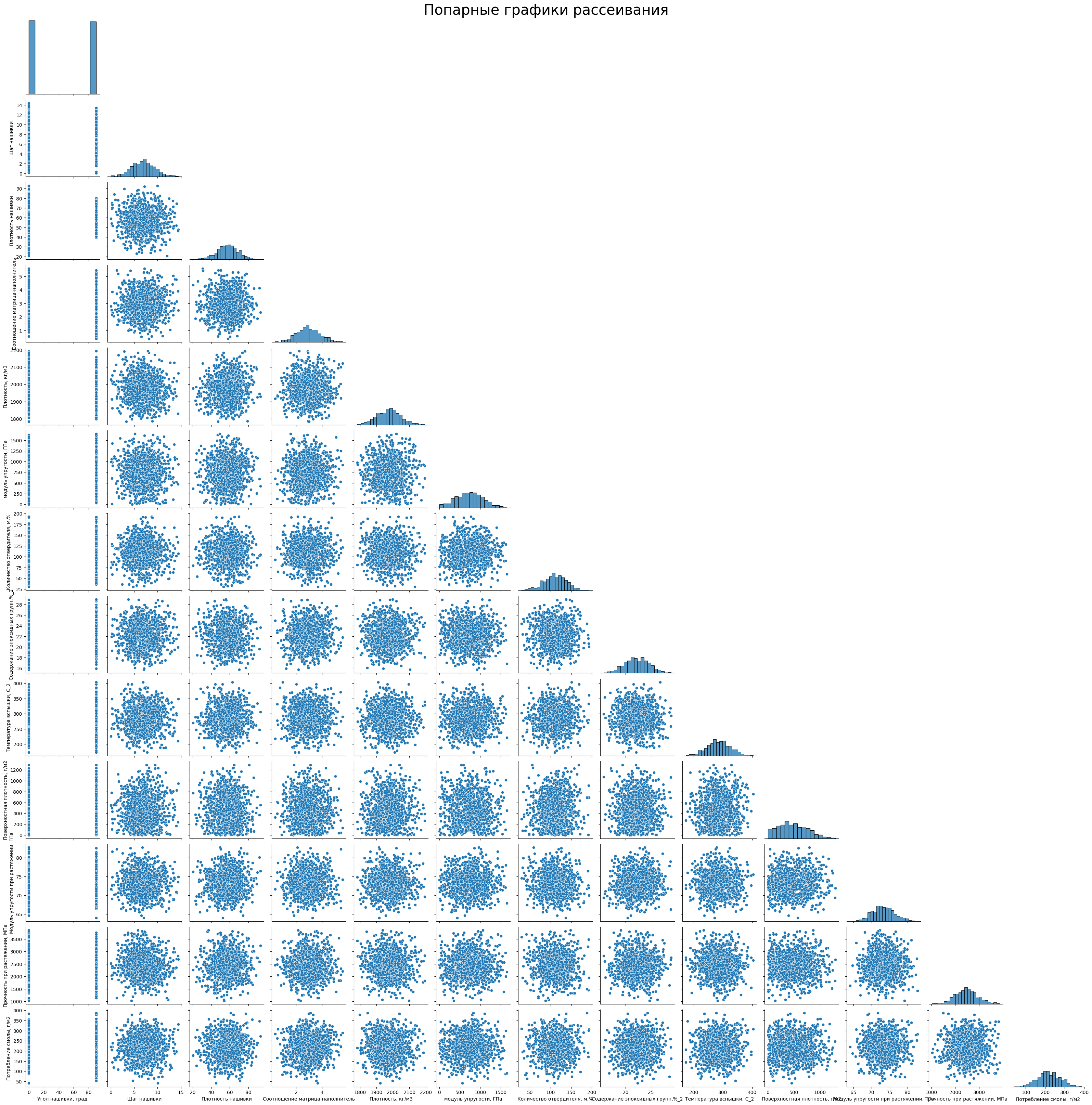




Датасет был предварительно обработан, что объясняет отсутствие пропусков. В исходных необработанных данных обычно встречаются пропуски и значения некорректных типов.

Попарные графики рассеяния точек, приведенные на рисунке 3, представляют взаимосвязи между признаками. На этих графиках видно, что некоторые точки значительно удалены от основного облака данных. Такие точки являются потенциальными выбросами — аномальными или некорректными значениями, которые выходят за пределы допустимых диапазонов признаков. График показывает, что некоторые точки стоят далеко от общего скопления.

Рисунок 3 – Попарный график рассеивания



Попробуем найти выбросы и подобрать для этой цели метод.

Метод Z-Score - xувствителен к выбросам, плохо работает с асимметричными данными. Хорошо работает, когда данные нормальны.

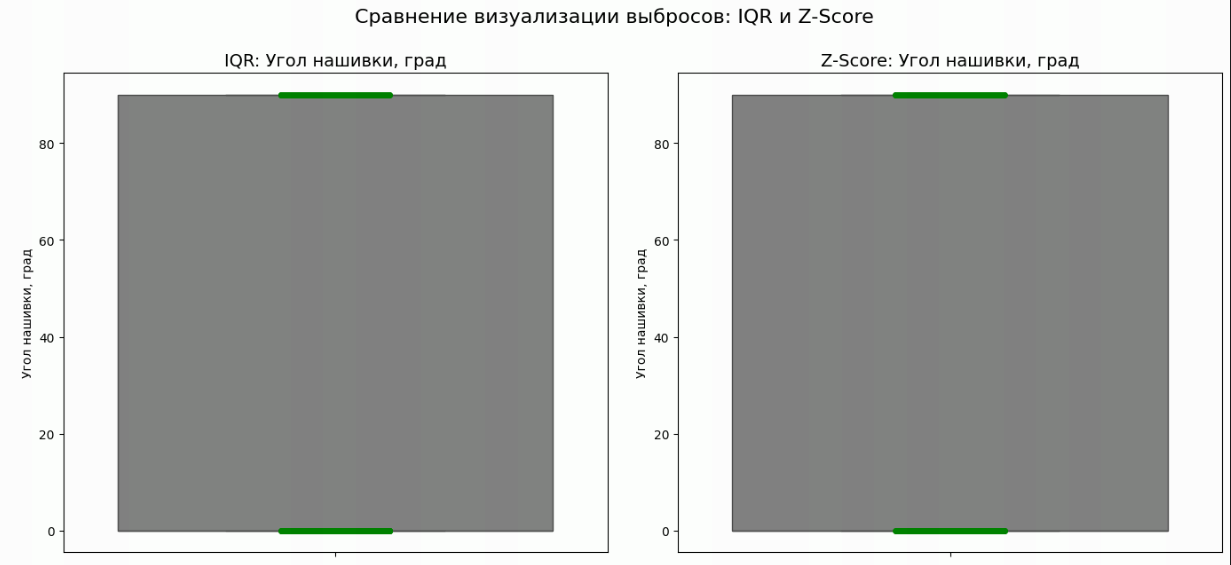
Метод IQR (межквартильного размаха) устойчив к выбросам, подходит для данных с неравномерным распределением. Не учитывает общую структуру распределения. Применим, когда данные асимметричны.

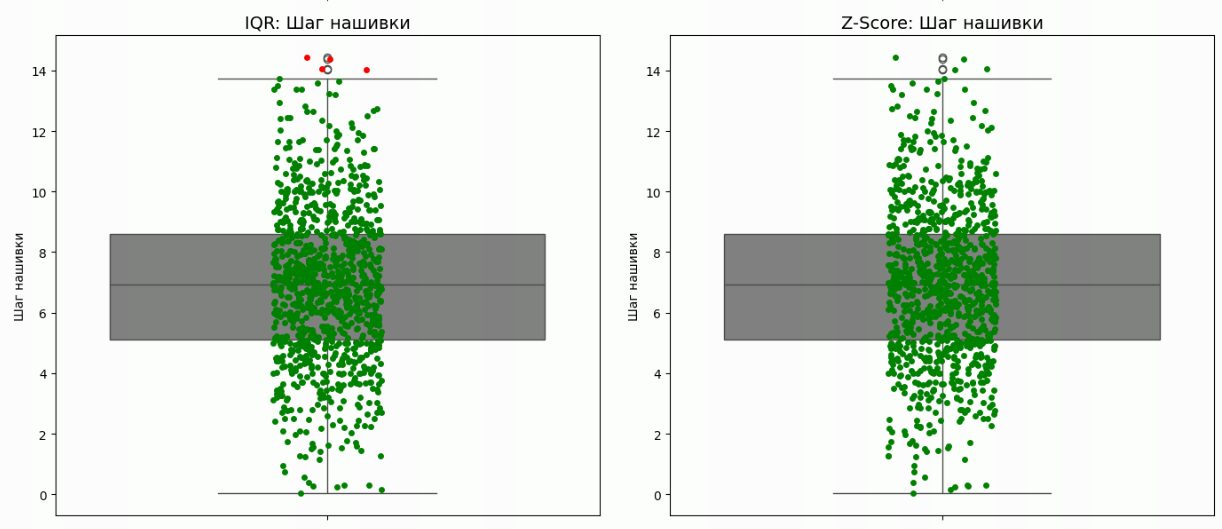
Метод DBSCAN - применим для сложных, многомерных данных.

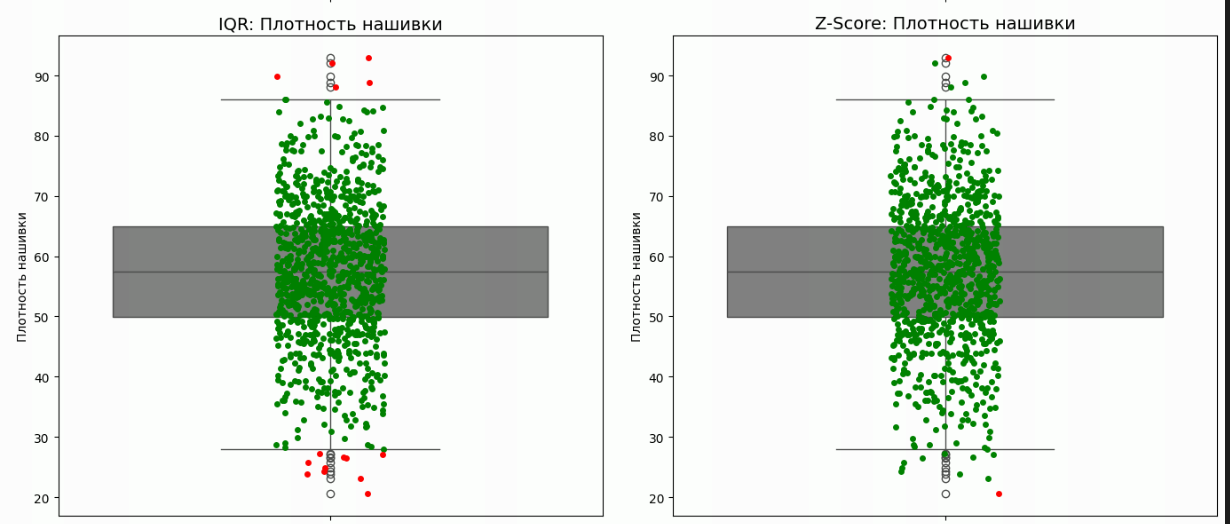
Метод пороговых значений - Требует ручной настройки, подходит не для всех данных.

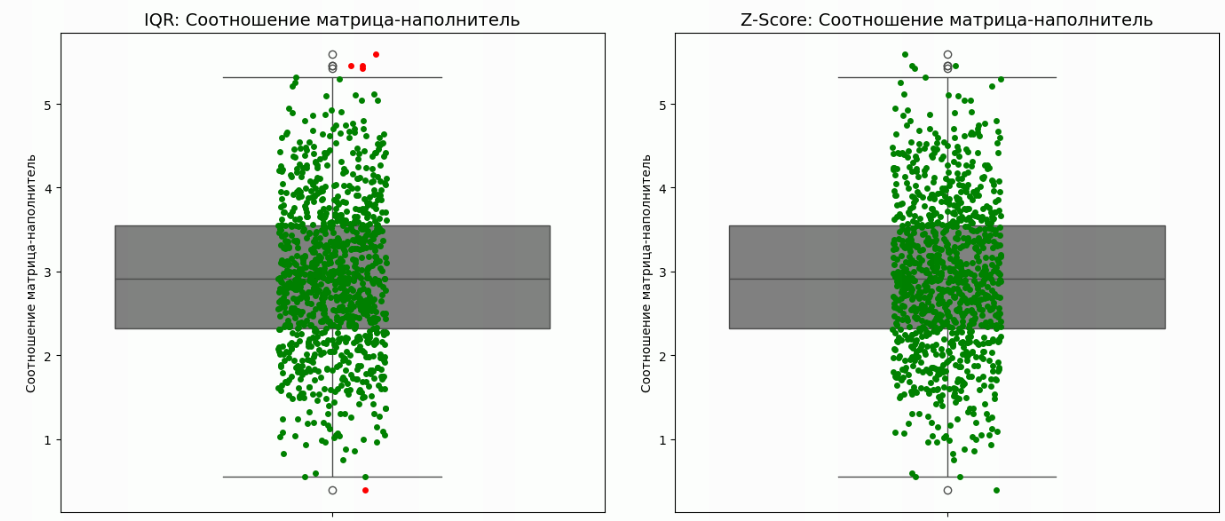
Так как наши данные нормальны, используем метод Z-Score, и сравним с методом IQR. Мы используем метод Z-Score, так как наши данные нормальны, но сравним их с методом IQR. График представлен на рисунке 4.

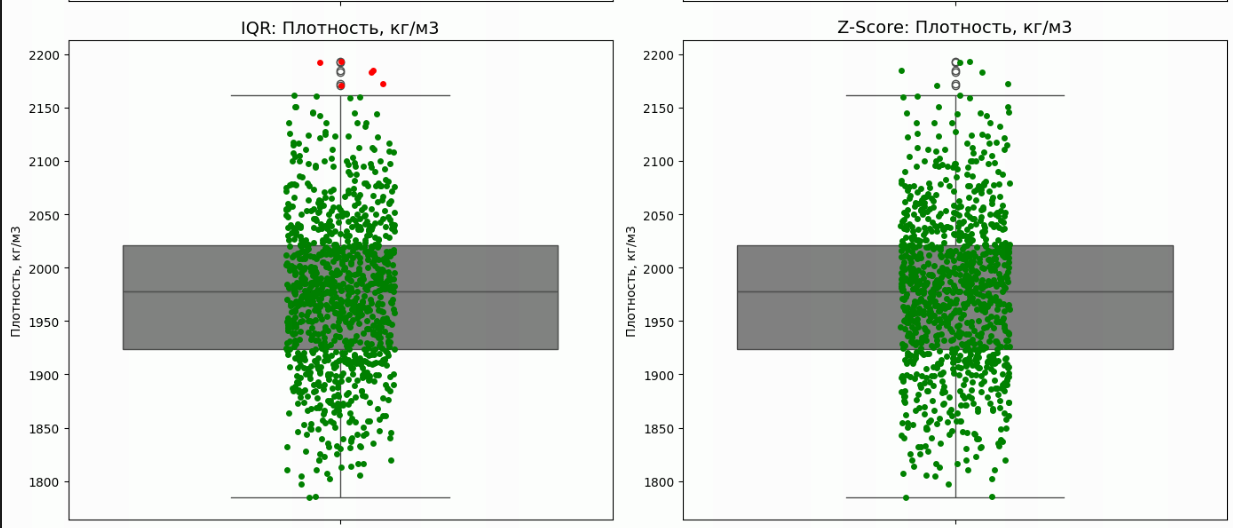
Рисунок 4 – Визуализация выбросов

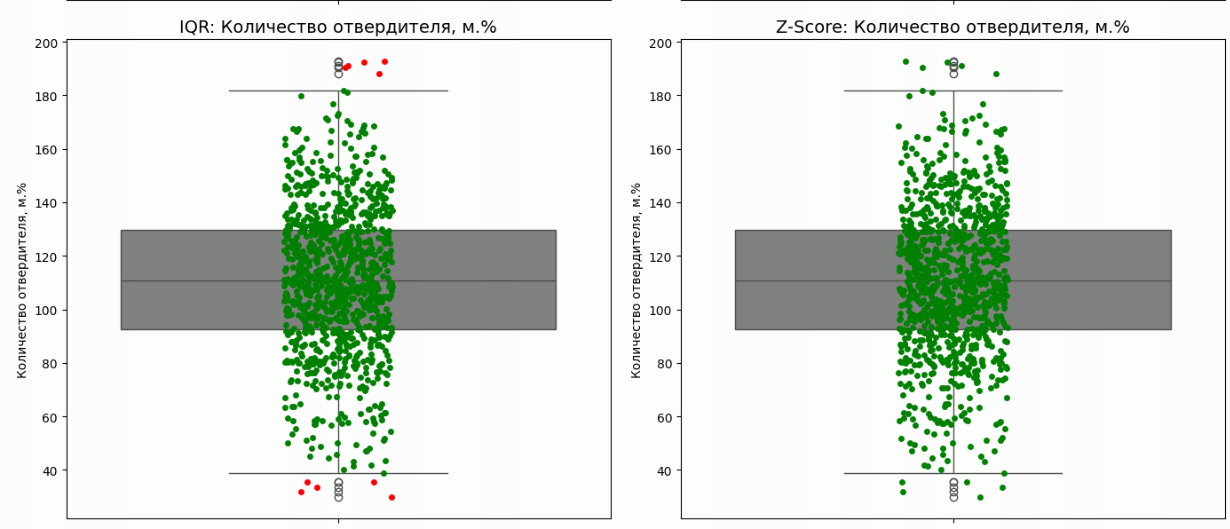
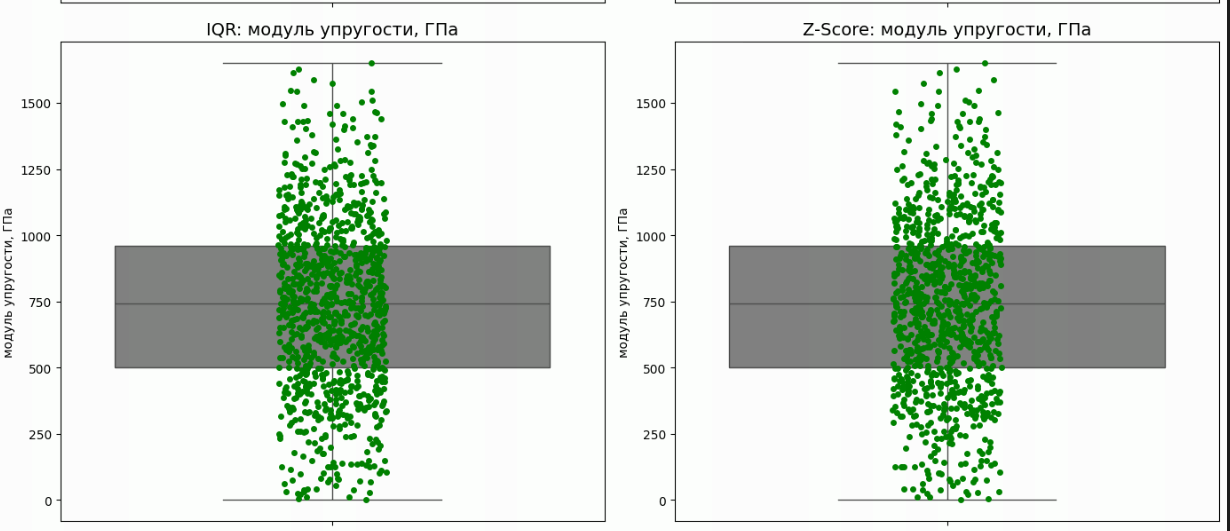


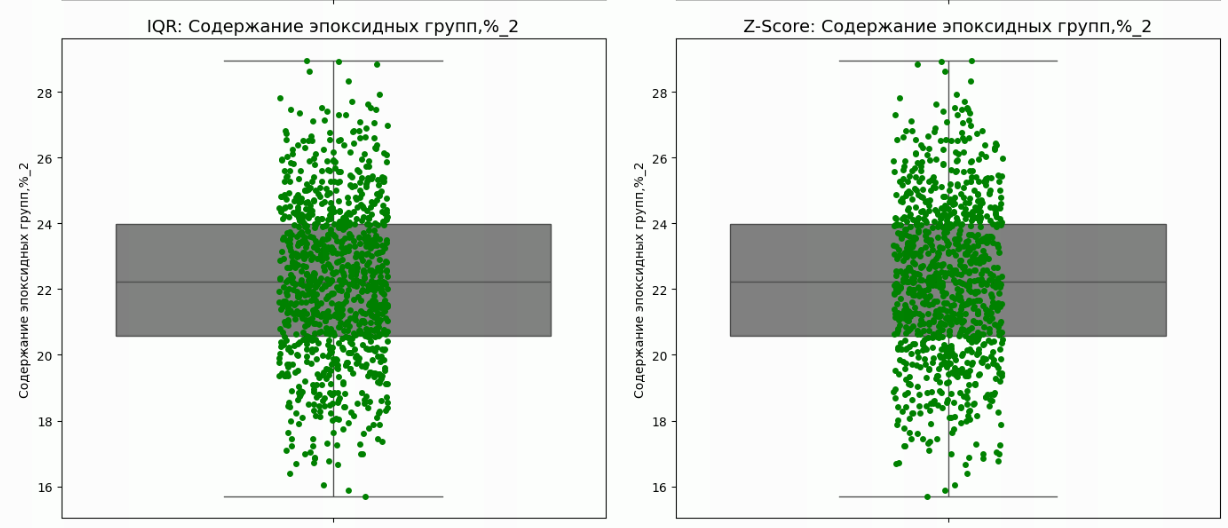


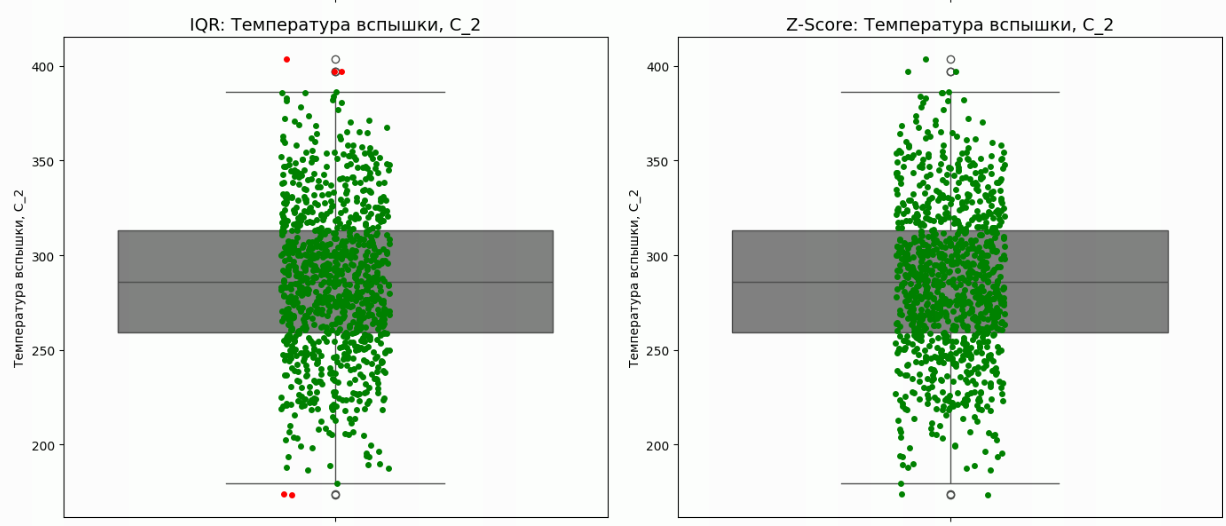


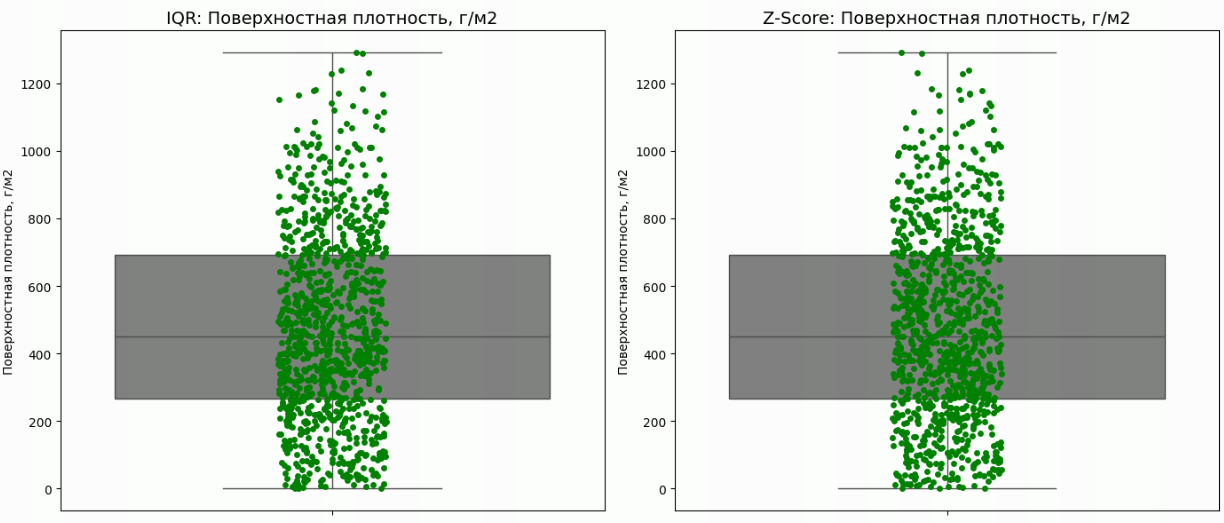


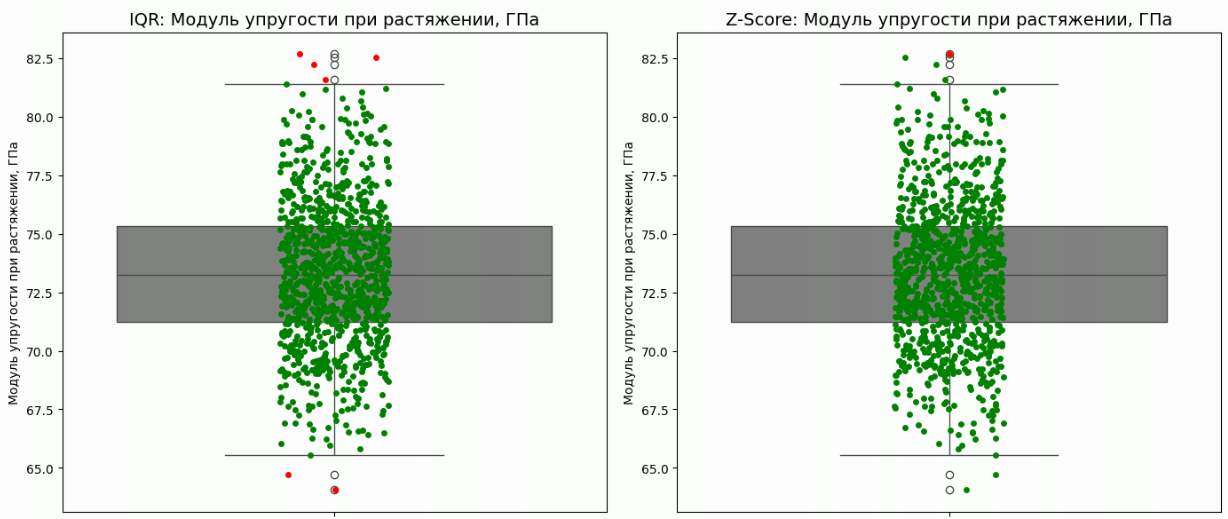


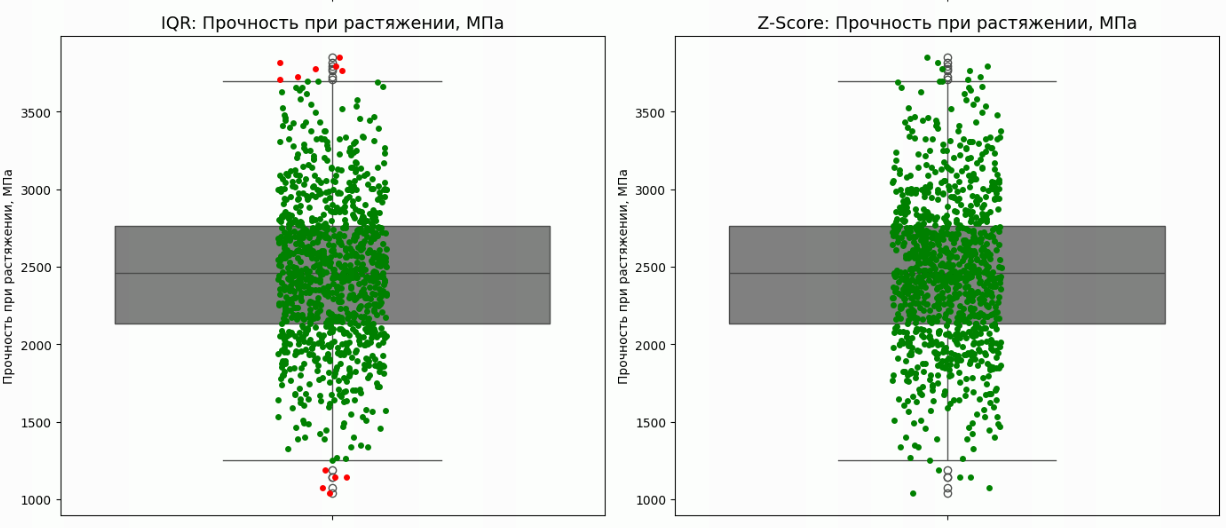


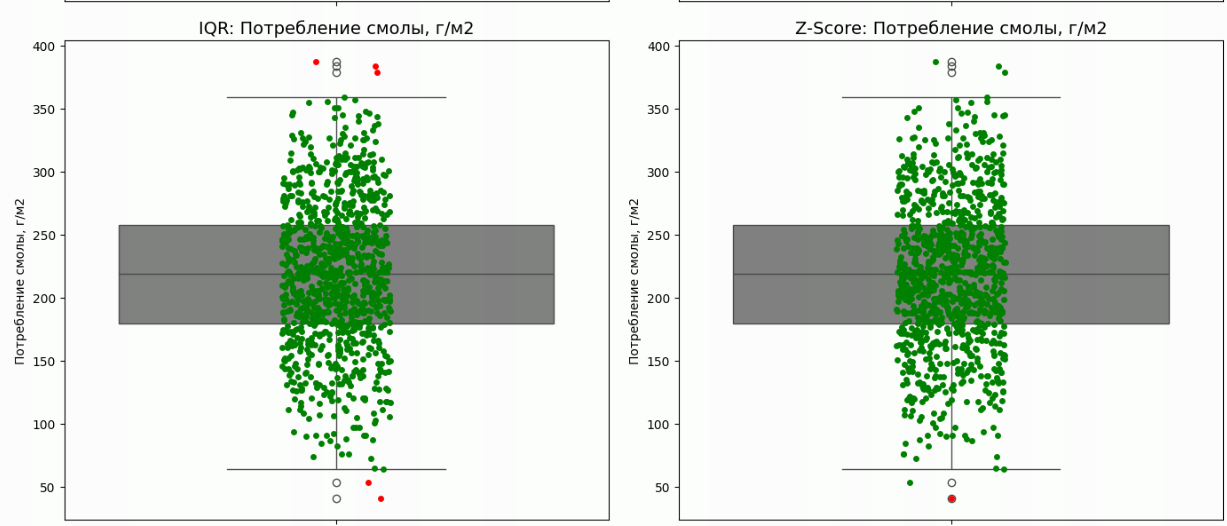








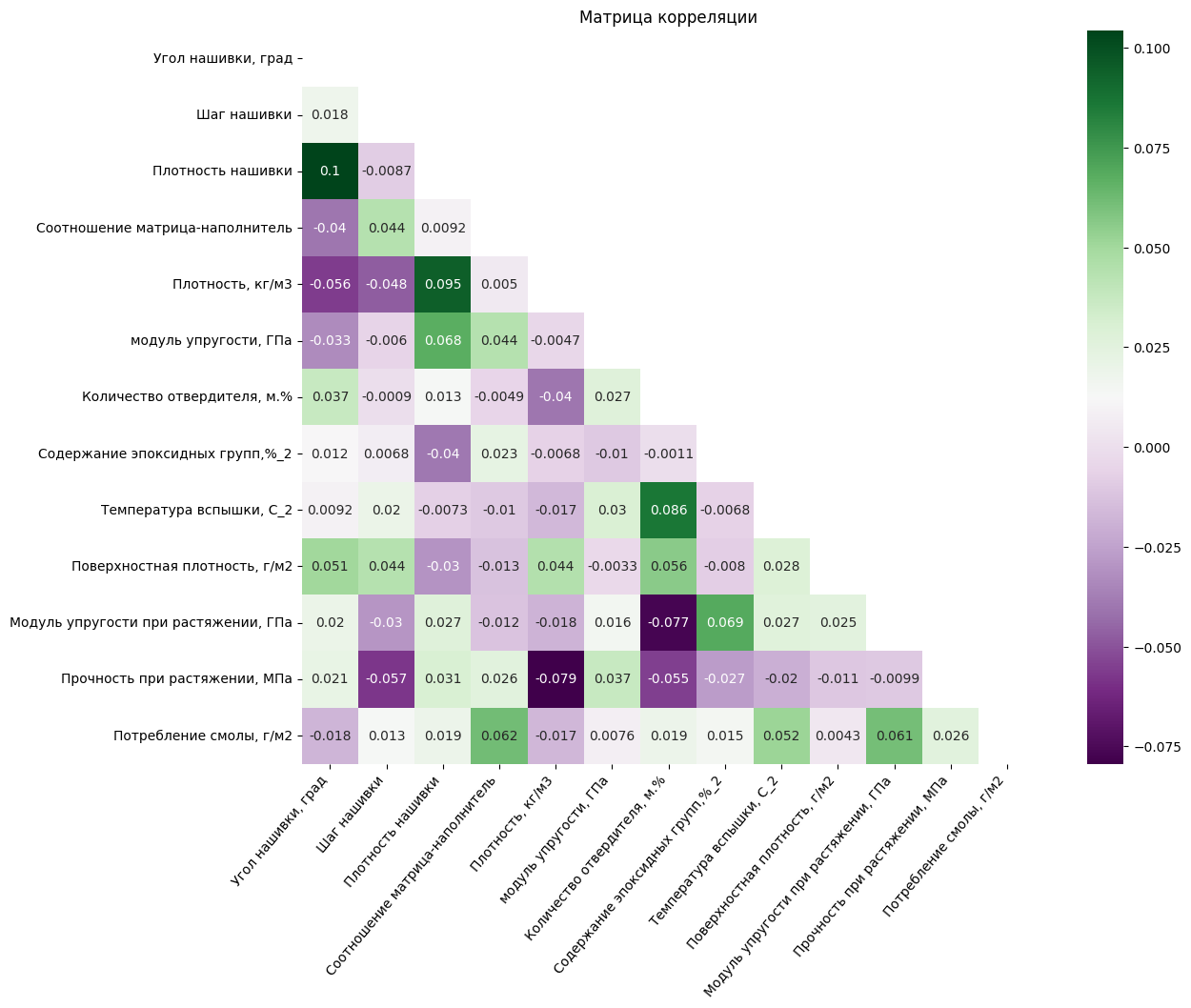




Построим корреляционную матрицу, приведена на рисунке 5. Для определения взаимосвязи между переменными. Анализ взаимосвязей помогает понять, насколько сильно и в каком направлении связаны переменные. Это важно при построении моделей, чтобы избежать избыточности данных, которые дают высокие корреляции между независимыми переменными и могут ухудшить работу моделей, особенно линейных. Если две переменные сильно коррелируют, одну из них можно исключить из модели для упрощения анализа. Также, матрица может выявить интересные взаимосвязи, которые требуют дополнительного изучения.

Но необходимо учитывать, что коэффициенты корреляции показывают только линейные зависимости. Если связь нелинейная, она может быть не выявлена, а также корреляция не указывает на причинно-следственные связи. Высокая корреляция не означает, что одна переменная влияет на другую.

Рисунок 5 – Корреляционная матрица в виде тепловой карты



Из тепловой карты находим, что максимальная корреляция между плотностью нашивки и углом нашивки равна 0,1, корреляция между остальными параметрами близка к нулю. Это значит, что линейные зависимости между ними отсутствуют. И линейная модель регрессии не даст приемлемого результата.

# 1.2 Описание используемых методов

Поскольку линейные зависимости не найдены, рассмотрим другие методы регрессии. Прогнозирование значений непрерывной числовой переменной относится к задаче регрессии, зависимая переменная связана с одной или несколькими независимыми переменными, которые также называют предикторами или регрессорами. Регрессионный анализ позволяет изучить, как среднее значение зависимой переменной изменяется в зависимости от изменений независимых переменных.

Нормализуем данные методом MinMaxScaler — это метод масштабирования данных, который преобразует значения признаков таким образом, чтобы они попадали в заданный диапазон, обычно от нуля до единицы.

Рассмотрим выбранные модели машинного обучения.

# 1.2.1 Ансамбль из множества деревьев решений

RandomForestRegressor — это метод машинного обучения, который использует ансамбль из множества деревьев для решения задачи регрессии (предсказания числовых значений). Каждый из этих деревьев делает своё предсказание, а итоговый результат вычисляется как среднее значение всех деревьев.

1. Данные делятся на множество небольших частей.

2. Для каждой части строится своё дерево решений.

3. Все деревья "голосуют" за предсказание, и берётся их среднее значение.

Работа метода: данные рандомно делятся на обучающие наборы для каждого дерева. Каждое дерево строится независимо от других. Деревья обучаются на случайных подмножествах признаков, чтобы избежать переобучения и сделать модель более устойчивой. На этапе предсказания все деревья делают свои прогнозы, а результат усредняется.

Плюсы метода:

1. Высокая точность: метод часто работает лучше, чем отдельное дерево решений, так как сглаживает ошибки отдельных деревьев.
2. Устойчивость к переобучению: благодаря усреднению, модель менее склонна запоминать шумы в данных.
3. Обработка нелинейных зависимостей: хорошо работает даже с сложными данными.
4. Может работать с разнородными данными: хорошо справляется как с числовыми, так и с категориальными признаками.
5. Простота настройки: не требует много тонкой настройки гиперпараметров.
6. Оценка важности признаков: можно увидеть, какие признаки вносят наибольший вклад в предсказание.

Минусы метода:

1. Медленная работа на больших данных: построение множества деревьев требует времени и ресурсов, особенно для больших наборов данных.
2. Сложность интерпретации: итоговый результат создаётся на основе множества деревьев, что делает модель менее интерпретируемой, чем простые методы (например, линейная регрессия).
3. Неэффективность на высокоразмерных данных: если у данных очень много признаков, модель может становиться менее точной из-за случайного выбора признаков для построения деревьев.
4. Большое потребление памяти: хранение большого количества деревьев требует значительных ресурсов.

Данный метод хорошо подходит для сложных задач регрессии, где линейные модели не дают хороших результатов и для анализа данных, где важна оценка приоритетности признаков.

# 1.2.2 Метод K-ближайших соседей (K-Nearest Neighbors, KNN)

Это алгоритм машинного обучения, который используется для решения задач регрессии (предсказания числовых значений). Он основывается на идее, что похожие объекты должны иметь схожие результаты

Принцип работы достаточно: для каждой новой точки, которую нужно предсказать, алгоритм находит K ближайших соседей в обучающих данных. После этого делает предсказание, усредняя значения целевой переменной этих соседей (или используя другие способы, такие как взвешенное среднее).

Плюсами данного метода являются простота, KNeighborsRegressor легко понять и реализовать, он не делает предположений о форме данных, таких как линейность или нормальность, алгоритм может работать с различными метриками расстояния (например, евклидово расстояние, манхэттенское расстояние).

Минусы метода. Для больших данных или многомерных данных вычисления расстояний для каждого предсказания могут быть очень медленными. Результаты сильно зависят от выбранного числа соседей (K). Если K слишком маленькое, модель может быть чувствительна к шуму. Если K слишком большое, она может быть слишком сглаженной и упускать важные детали. В высоких измерениях (многомерные данные) становится сложнее найти «ближайших соседей», и метод может терять свою эффективность из-за так называемого парадокса измерения.

# 1.2.3 Метод Gradient Boosting

GradientBoostingRegressor — это алгоритм машинного обучения для решения задач регрессии, основанный на деревьях решений. Он работает путем построения ансамбля слабых моделей (обычно деревьев решений), которые объединяются, чтобы дать сильное предсказание.

Рассмотрим принцип работы метода.

Алгоритм начинает с простой модели, например, предсказывая среднее значение целевой переменной. На каждом шаге алгоритм вычисляет ошибки (остатки) текущей модели. Строится новое дерево решений, которое учится предсказывать ошибки предыдущей модели. Новое дерево добавляется к ансамблю, чтобы уменьшить ошибки. Этот процесс повторяется заданное число раз (или до достижения нужной точности). Итоговое предсказание — это сумма предсказаний всех деревьев с учетом их весов.

Плюсами метода являются высокая точность, благодаря своей способности хорошо улавливать сложные зависимости в данных. Он подходит для работы с различными типами данных и не требует их масштабирования. Использование регуляризации (например, ограничение глубины деревьев, темп обучения) позволяет избежать переобучения.

Минусы метода сводятся к построению большого числа деревьев, что может замедлять его работу, особенно на больших наборах данных. Результат обработки данных сильно зависит от таких параметров, как глубина деревьев, число деревьев, темп обучения. Самый большой минус модели в том, что он чувствителен к шуму и выбросам, поэтому данные перед использованием должны быть тщательно обработаны.

GradientBoostingRegressor подходит для задач, где требуется высокая точность, и есть время на настройку гиперпараметров. Если данные сложные или имеют нелинейные зависимости, этот метод может стать отличным выбором.

# 1.2.4 Метод с поддерживающими векторами (Support Vector Regression, SVR)

Это алгоритм машинного обучения, основанный на методе опорных векторов (SVM), который применяется для решения задач регрессии.

Принцип работы метода заключается в том, что SVR пытается найти гиперплоскость (в многомерном пространстве), которая максимально точно описывает зависимость между входными и выходными данными. При этом допускаются небольшие ошибки (в пределах заданного порога ε) для большей гибкости. Вместо минимизации ошибок для всех точек SVR фокусируется только на ключевых точках данных (опорных векторах), которые определяют положение гиперплоскости.

К плюсам метода можно отнести гибкость, он может моделировать как линейные, так и нелинейные зависимости (с помощью ядер, например, RBF или полиномиальных), робустность, т.е. устойчивость к выбросам, так как они не оказывают большого влияния на гиперплоскость, и контроль точности, так как метод настраиваемый и в нем можно задавать допустимую погрешность через параметр ε.

К минусам метода можно отнести высокую вычислительную сложность, так как он требует больших ресурсов на больших или высокоразмерных данных. Так же он чувствительность к гиперпараметрам несмотря на то, что в нем можно тщательно настраивать параметры, такие как C, ε, и тип ядра. Так же модель менее интуитивно понятна, чем простые методы.

SVR хорошо подходит для задач с небольшим числом признаков и выборкой, особенно если есть сложные зависимости между данными, и требуется высокая точность.

# 1.3 Разведочный анализ данных

Разведочный анализ данных представляет собой важный этап при работе с данными, поскольку его цель заключается в выявлении скрытых закономерностей и взаимосвязей внутри набора данных. Этот процесс помогает лучше понять структуру данных и определить, какие признаки могут оказывать влияние на результаты модели.

При построении большинства моделей машинного обучения требуется выполнение двух ключевых условий:

1. Зависимость выходных переменных от входных: чем сильнее эта зависимость, тем точнее модель сможет предсказывать результат на основе входных признаков.

2. Независимость между самими входными переменными: наличие сильной корреляции между входными признаками может привести к переобучению модели и снижению её точности.

Ранее, на примере графика попарного рассеяния точек (рисунок 3), мы могли наблюдать распределение данных. Визуальный анализ показал, что форма «облаков точек» не позволяет сразу заметить явные зависимости, которые были бы полезны для построения моделей. Однако визуализация — лишь первый шаг в анализе данных. Для выявления более тонких связей между признаками часто используется такой инструмент, как матрица корреляции.

Матрица корреляции, представленная ранее на рисунке 5, даёт количественную оценку степени линейной связи между различными признаками. Это позволяет оценить, насколько сильно одни признаки влияют на другие, а также обнаружить возможные мультиколлинеарности, когда несколько признаков оказываются тесно связаны друг с другом. Такой подход помогает сделать вывод о том, какие признаки стоит включить в модель, а какие, возможно, следует исключить для повышения качества прогнозирования.

Таким образом, разведочный анализ данных играет ключевую роль в подготовке данных для дальнейшего моделирования, позволяя исследователю получить представление об основных характеристиках и взаимосвязях в наборе данных.

# 1.3.1 Выбор признаков

Анализ корреляции показал, что все коэффициенты корреляции между признаками близки к нулю, что указывает на отсутствие линейной зависимости между ними. Это может свидетельствовать о том, что признаки не связаны напрямую или существуют сложные нелинейные зависимости, которые невозможно выявить простыми статистическими методами. Важно подчеркнуть, что не всегда низкая корреляция означает отсутствие взаимосвязи, так как могут быть скрытые зависимости, которые требуют более сложных подходов, например, методов машинного обучения. В контексте этого исследования мы можем предположить, что признаки делятся на несколько групп:

Свойства матрицы - характеристики материала, такие как модуль упругости, плотность и другие, которые определяют основные физико-химические свойства.

2. Свойства наполнителя - параметры, которые влияют на прочность и другие механические характеристики композита.

3. Свойства смеси и производственного процесса - параметры, связанные с технологией создания композита, такие как температура, время отверждения и другие условия.

4. Свойства готового композита - итоговые характеристики готового продукта, такие как прочность, эластичность и другие эксплуатационные качества.

Таким образом, группы признаков могут влиять на конечные характеристики композита, и каждый из этих факторов стоит учитывать при построении модели. Также важно отметить, что точное распределение признаков по этим категориям может требовать уточнения на основе более глубокого знания предметной области.

На основе зависимостей будут построены отдельные модели для каждого целевого признака, что позволяет решать три независимые задачи.

# 1.3.2 Препроцессинг

Препроцессинг данных играет ключевую роль в обеспечении корректной работы моделей. Важно отметить, что предварительная обработка должна проводиться после разделения данных на тренировочную и тестовую выборки, чтобы избежать утечек информации.

Категориальный признак в этом исследовании — "Угол нашивки, град", который принимает значения 0 и 90.

Для вещественных признаков возникает проблема различия в диапазонах значений, что может привести к искажению результатов при обучении модели.

Для исправления, применяются методы нормализации или стандартизации. В данном случае выбрана стандартизация, поскольку она более предпочтительна для алгоритмов, чувствительных к масштабу, таких как линейные модели или методы на основе градиентного спуска. Стандартизация приводит признаки к нулевому среднему и единичному стандартному отклонению, что помогает моделям работать более эффективно.

# 1.3.3 Поиск гиперпараметров по сетке

Поиск гиперпараметров по сетке (GridSearchCV) позволяет автоматически оптимизировать параметры модели, пробуя различные их комбинации и оценивая результат с помощью перекрестной проверки. Это помогает найти наилучшие параметры, которые улучшают производительность модели.

# 1.3.4 Метрики качества моделей

Для оценки эффективности моделей в задаче регрессии используются следующие метрики:

- R2 (коэффициент детерминации): показывает, какая доля дисперсии целевой переменной объясняется моделью. Чем ближе значение R2 к единице, тем лучше модель объясняет данные. Если R2 близок к нулю, модель не объясняет данные и ее прогнозы хуже, чем простое предсказание среднего значения.

- MSE (Mean Squared Error) — это метод оценки ошибки модели, который вычисляет среднее значение квадратов разностей между реальными значениями и предсказанными моделью. Чем меньше MSE, тем точнее модель.

- MAE (средняя абсолютная ошибка): метрика, которая оценивает среднюю абсолютную разницу между предсказанными и истинными значениями. Она более устойчива к выбросам по сравнению с MSE.

# 2 Практическая часть

# 2.1. Разбиение и предобработка данных

# 2.1.1 Для прогнозирования модуля упругости при растяжении

Признаки датасета были разделены на входные и выходные, а строки – на тренировочное и тестовое множество. Размерности полученных наборов данных показаны на рисунке 6. Описательная статистика входных признаков после предобработки показана на рисунке 7.

x1\_train: (700, 12) y1\_train: (700,)

x1\_test: (300, 12) y1\_test: (300,)

Рисунок 6 – размерности набора данных



Рисунок 7 – Описательная статистика признаков

# 2.1.2 Для прогнозирования прочности при растяжении

Признаки датасета были разделены на входные и выходные, а строки – на тренировочное и тестовое множество. Размерности полученных наборов данных показаны на рисунке 8.

Рисунок 8 – размерности набора данных

2.2 Разработка и обучение моделей для прогнозирования модуля упругости при растяжении и модуля прочности при растяжении.

Для подбора лучшей модели для этой задачи были подобраны модели, описанные в разделе 1.2.

Метрики работы выбранных моделей с гиперпараметрами по умолчанию, полученные с помощью перекрестной проверки на тестовом множестве, приведены на рисунке 9.

Рисунок 9 – Сравнение моделей базовых моделей

Ни одна из выбранных мной моделей не оказалась подходящей для наших данных. Коэффициент детерминации R2 близок к нулю. Значит, они не лучше базовой модели. И остальные метрики у них примерно совпадают с базовой моделью.

После выполнения подбора гиперпарметров по сетке с перекрестной проверкой, можно сделать вывод, что, подбирая гиперпараметры, можно значительно улучшить предсказание выбранной модели.

Все модели не очень хорошо описывают исходные данные - не удалось добиться положительного значения R2. Самая лучшая модель дает коэффициент детерминации близкий к нулю, что соответствует базовой модели.

# Заключение

В ходе выполнения данной работы были пройдены основные этапы Dataflow pipeline. Были применены теоретические методы анализа данных и машинного обучения.

Процесс включал несколько ключевых этапов:

1. Изучение теоретических методов анализа данных и машинного обучения, это основа для выбора правильных алгоритмов и моделей.

2. Ознакомление с предметной областью, для интерпретации данных

3. Извлечение и трансформация данных. Готовый набор данных упростил этап.

4. Проведение разведочного анализа данных (EDA), который дал представление о характере данных и возможных зависимостях между признаками.

5. Выполнение предобработки данных для обеспечения корректной работы моделей, включая стандартизацию и обработку категориальных признаков.

6. Построение аналитического решения, выбор и оценка различных моделей, а также подбор гиперпараметров для достижения наилучших результатов.

Работа с нейросетями требует значительных усилий и глубокого понимания принципов их функционирования. Для построения модели на основе нейросетей требуется более глубокое погружение в предмет и большая практика работы.

Библиографический список

1. Иванов Д. А., Ситников А. И., Шляпин С. Д. Композиционные материалы: учебное пособие для вузов / под ред. А. А. Ильина. — Москва: Издательство Юрайт, 2019. — 253 с. — (Высшее образование). — Текст: непосредственный.

2. Силен Д., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных. — СПб.: Питер, 2017. — 336 с.: ил.

3. Грас Д. Data Science. Наука о данных с нуля: Пер. с англ. — 2-е изд., перераб. и доп. — СПб.: БХВ-Петербург, 2021. — 416 с.: ил.

4. Документация по языку программирования Python. — Режим доступа: https://docs.python.org/3.11/index.html.

5. Документация по библиотеке NumPy. Режим доступа: https://numpy.org/doc/1.23/user/index.html#user.

6. Документация по библиотеке Pandas. Режим доступа: https://pandas.pydata.org/docs/user\_guide/index.html#user-guide.

7. Документация по библиотеке Matplotlib. Режим доступа: https://matplotlib.org/stable/users/index.html.

8. Документация по библиотеке Seaborn. Режим доступа: https://seaborn.pydata.org/tutorial.html.

9. Документация по библиотеке Scikit-learn. Режим доступа: https://scikit-learn.org/stable/user\_guide.html.

10. Документация по библиотеке Keras. Режим доступа: https://keras.io/api/.

11. Руководство по быстрому старту в Flask. Режим доступа: https://flaskrussian-docs.readthedocs.io/ru/latest/quickstart.html.

12. Loginom Вики. Алгоритмы. Режим доступа: https://wiki.loginom.ru/algorithms.html.

13. Maszański, A. Метод k-ближайших соседей (k-nearest neighbour). — Режим доступа: https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearestneighbour-2021-07-19.

14. Kashnitsky, Y. Открытый курс машинного обучения. Тема 3. Классификация, деревья решений и метод ближайших соседей. — Режим доступа: https://habr.com/ru/company/ods/blog/322534/.

15. Maszański, A. Машинное обучение для начинающих: алгоритм случайного леса (Random Forest). — Режим доступа: https://proglib.io/p/mashinnoeobuchenie-dlya-nachinayushchih-algoritm-sluchaynogo-lesa-random-forest-2021-08-12.

16. Maszański, A. Решаем задачи машинного обучения с помощью алгоритма градиентного бустинга. — Режим доступа: https://proglib.io/p/reshaem-zadachi-mashinnogo-obucheniya-s-pomoshchyu-algoritma-gradientnogo-bustinga-2021-11-25.