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»

Слушатель Колесников Анатолий Федорович (ФИО)

Москва, 2023

# Оглавление

[Введение 3](#_Toc106468218)

[**1. Аналитическая часть** 5](#_Toc106468219)

[**2. Практическая часть** 24](#_Toc106468220)

[Заключение 28](#_Toc106468221)

# **Введение**

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Яркий пример композита - железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

**На входе** имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

**Актуальность:** Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

# **1. Аналитическая часть**

**1.1. Постановка задачи**

Композиционный материал - неоднородный сплошной материал, состоящий из двух или более компонентов, среди которых можно выделить армирующие элементы, обеспечивающие необходимые механические характеристики материала, и матрицу (или связующее), обеспечивающую совместную работу армирующих элементов.

Композиты - многокомпонентные материалы, состоящие из полимерной, металлической., углеродной, керамической или др. основы (матрицы), армированной наполнителями из волокон, нитевидных кристаллов, тонкодиспeрсных частиц и др. Путем подбора состава и свойств наполнителя и матрицы (связующего), их соотношения, ориентации наполнителя можно получить материалы с требуемым сочетанием эксплуатационных и технологических свойств. Использование в одном материале нескольких матриц (полиматричные композиционные материалы) или наполнителей различной природы (гибридные композиционные материалы) значительно расширяет возможности регулирования свойств композиционных материалов.

Эффективность и работоспособность материала зависят от правильного выбора исходных компонентов и технологии их совмещения, призванной обеспечить прочную связь между компонентами при сохранении их первоначальных характеристик. В результате совмещения армирующих элементов и матрицы образуется комплекс свойств композита, не только отражающий исходные характеристики его компонентов, но и включающий свойства, которыми изолированные компоненты не обладают. В частности, наличие границ раздела между армирующими элементами и матрицей существенно повышает трещиностойкость материала, и в композитах, в отличие от металлов, повышение статической прочности приводит не к снижению, а, как правило, к повышению характеристик вязкости разрушения.

Современные композиты изготавливаются из различных материалов, но даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из известных компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов (затратно по времени, и потребляемым ресурсам) или теоретическое прогнозирование характеристик. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

Прогнозирование модели может существенно сократить количество проводимых испытаний, а также пополнить базу данных материалов новыми свойствами материалов, и цифровыми двойниками новых композитов. Актуальность решения задачи обусловлена широким использованием композитных материалов практически во всех областях производства.

Предметом выпускной квалификационной работы является построение моделей для прогнозирования таких характеристик композиционных материалов, как модуль упругости при растяжении, прочность при растяжении и создание нейронной сети для рекомендации соотношения матрица-наполнитель.

Исходные датасеты о свойствах композиционных материалов получены МГТУ им. Н.Э. Баумана – Центр НТИ «Цифровое материаловедение: новые материалы и вещества».

Датасет состоит из двух файлов:

1. файл X\_bp.xlsx с данными о параметрах базальтопластика;
2. файл X\_nup.xlsx с данными о нашивках из углепластика.

Далее проводится объединение двух файлов X\_bp.xlsx и X\_nup.xlsx по индексу, используя тип объединения INNER.

Количество строк в файле X\_bp.xlsx было 1023, столбцов 10. А количество строк в файле X\_nup.xlsx – 1040, столбцов 3.

После объединения таблиц, 17 строк из файла X\_nup.xlsx было отброшено. Дальнейшие исследования проводились с датасетом содержащим 1023 строк и 13 столбцов.

Общая информация о таблице

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Column** | **Count** | **Non-Null** | **Dtype** |
| 0 | Соотношение матрица-наполнитель | 1023 | non-null | float64 |
| 1 | Плотность, кг/м3 | 1023 | non-null | float64 |
| 2 | модуль упругости, Гпа | 1023 | non-null | float64 |
| 3 | Количество отвердителя, м.% | 1023 | non-null | float64 |
| 4 | Содержание эпоксидных групп,%\_2 | 1023 | non-null | float64 |
| 5 | Температура вспышки, С\_2 | 1023 | non-null | float64 |
| 6 | Поверхностная плотность, г/м2 | 1023 | non-null | float64 |
| 7 | Модуль упругости при растяжении, Гпа | 1023 | non-null | float64 |
| 8 | Прочность при растяжении, Мпа | 1023 | non-null | float64 |
| 9 | Потребление смолы, г/м2 | 1023 | non-null | float64 |
| 10 | Угол нашивки, град | 1023 | non-null | int64 |
| 11 | Шаг нашивки | 1023 | non-null | float64 |
| 12 | Плотность нашивки | 1023 | non-null | float64 |

Описательная статистика

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Соотношение матрица-наполнитель** | 1023 | 2.930366 | 0.913222 | 0.389403 | 2.317887 | 2.906878 | 3.552660 | 5.591742 |
| **Плотность, кг/м3** | 1023 | 1975.734888 | 73.729231 | 1731.764635 | 1924.155467 | 1977.621657 | 2021.374375 | 2207.773481 |
| **модуль упругости, ГПа** | 1023 | 739.923233 | 330.231581 | 2.436909 | 500.047452 | 739.664328 | 961.812526 | 1911.536477 |
| **Количество отвердителя, м.%** | 1023 | 110.570769 | 28.295911 | 17.740275 | 92.443497 | 110.564840 | 129.730366 | 198.953207 |
| **Содержание эпоксидных групп,%\_2** | 1023 | 22.244390 | 2.406301 | 14.254985 | 20.608034 | 22.230744 | 23.961934 | 33.000000 |
| **Температура вспышки, С\_2** | 1023 | 285.882151 | 40.943260 | 100.000000 | 259.066528 | 285.896812 | 313.002106 | 413.273418 |
| **Поверхн. плотность, г/м2** | 1023 | 482.731833 | 281.314690 | 0.603740 | 266.816645 | 451.864365 | 693.225017 | 1399.542362 |
| **Модуль упругости при растяжении, ГПа** | 1023 | 73.328571 | 3.118983 | 64.054061 | 71.245018 | 73.268805 | 75.356612 | 82.682051 |
| **Прочность при растяжении, МПа** | 1023 | 2466.922843 | 485.628006 | 1036.856605 | 2135.850448 | 2459.524526 | 2767.193119 | 3848.436732 |
| **Потребление смолы, г/м2** | 1023 | 218.423144 | 59.735931 | 33.803026 | 179.627520 | 219.198882 | 257.481724 | 414.590628 |
| **Угол нашивки, град** | 1023 | 44.252199 | 45.015793 | 0.000000 | 0.000000 | 0.000000 | 90.000000 | 90.000000 |
| **Шаг нашивки** | 1023 | 6.899222 | 2.563467 | 0.000000 | 5.080033 | 6.916144 | 8.586293 | 14.440522 |
| **Плотность нашивки** | 1023 | 57.153929 | 12.350969 | 0.000000 | 49.799212 | 57.341920 | 64.944961 | 103.988901 |

Из таблицы виден большой разброс значений и, следовательно, необходимость нормализации, которая будет проведена в дальнейшем.

1.2 Описание используемых методов

Линейная регрессия

Регрессия k-ближайших соседей

RandomForest

Рекомендательная нейросеть для соотношения матрица-наполнитель

Машинное обучение ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) machine learning, ML) — класс методов [искусственного интеллекта](https://ru.wikipedia.org/wiki/%D0%98%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82), характерной чертой которых является не прямое решение задачи, а обучение за счёт применения решений множества сходных задач. Для построения таких методов используются средства [математической статистики](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0), [численных методов](https://ru.wikipedia.org/wiki/%D0%A7%D0%B8%D1%81%D0%BB%D0%B5%D0%BD%D0%BD%D1%8B%D0%B5_%D0%BC%D0%B5%D1%82%D0%BE%D0%B4%D1%8B), [математического анализа](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7), [методов оптимизации](https://ru.wikipedia.org/wiki/%D0%9E%D0%BF%D1%82%D0%B8%D0%BC%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F_(%D0%BC%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0)), [теории вероятностей](https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BE%D1%80%D0%B8%D1%8F_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%B5%D0%B9), [теории графов](https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BE%D1%80%D0%B8%D1%8F_%D0%B3%D1%80%D0%B0%D1%84%D0%BE%D0%B2), различные техники работы с [данными в цифровой форме](https://ru.wikipedia.org/wiki/%D0%94%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5_(%D0%B2%D1%8B%D1%87%D0%B8%D1%81%D0%BB%D0%B8%D1%82%D0%B5%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F_%D1%82%D0%B5%D1%85%D0%BD%D0%B8%D0%BA%D0%B0)).

Машинное обучение — создание алгоритмов (набора правил), которые учатся на сложных функциях (шаблонах) из данных и делает прогнозы.  
Машинное обучение разделяется на несколько основных

подходов:

● Обучение с учителем (supervised learning)

наиболее распространенный и изученный тип машинного обучения, ведь легче обучить машину с выбранными данными. В зависимости от того, что вы хотите предсказать, обучение с учителем может использоваться для решения двух типов задач:

Классификация (classification)

Регрессия (regression)

Ранжирование (learning to rank)

● Обучение без учителя (unsupervised learning)

Поскольку у машин нет отмеченных данных для обучения, цель машинного обучения без учителя — обнаружение закономерностей в данных и их группировка.  
Обучение без учителя решает 2 типа задач:

Кластеризация (clustering)

Уменьшение размерности (dimensionality reduction)

● Обучение с частичным привлечением учителя

(semi-supervised learning)

Этот тип машинного обучения требует использования системы вознаграждения/штрафа. Цель — вознаградить машину, когда она учится правильно, и наказать машину, когда она учится неправильно.

● Обучение с подкреплением (reinforcement learning)

способ машинного обучения, при котором система обучается, взаимодействуя с некоторой средой.

**Типы задач машинного обучения**

Все задачи, решаемые с помощью ML, относятся к одной из следующих категорий.  
 **Задача регрессии**– прогноз на основе выборки объектов с различными признаками. На выходе должно получиться вещественное число (2, 35, 76.454 и др.), к примеру цена квартиры, стоимость ценной бумаги по прошествии полугода, ожидаемый доход магазина на следующий месяц, качество вина при слепом тестировании.

**Задача классификации** – получение категориального ответа на основе набора признаков. Имеет конечное количество ответов (как правило, в формате «да» или «нет»): есть ли на фотографии кот, является ли изображение человеческим лицом, болен ли пациент раком.

**Задача кластеризации** – распределение данных на группы: разделение всех клиентов мобильного оператора по уровню платёжеспособности, отнесение космических объектов к той или иной категории (планета, звёзда, чёрная дыра и т.п.).

**Задача уменьшения размерности** – сведение большого числа признаков к меньшему (обычно 2–3) для удобства их последующей визуализации(например, сжатие данных).

**Задача выявления аномалий** – отделение аномалий от стандартных случаев. На первый взгляд она совпадает с задачей классификации, но есть одно существенное отличие: аномалии – явление редкое, и обучающих примеров, на которых можно натаскать машинно обучающуюся модель на выявление таких объектов, либо исчезающе мало, либо просто нет, поэтому методы классификации здесь не работают. На практике такой задачей является, например, выявление мошеннических действий с банковскими картами.  
Основные алгоритмы моделей машинного обучения

**Дерево принятия решений**

Это метод поддержки принятия решений, основанный на использовании древовидного графа: модели принятия решений, которая учитывает их потенциальные последствия (с расчётом вероятности наступления того или иного события), эффективность, ресурсозатратность.

Для бизнес-процессов это дерево складывается из минимального числа вопросов, предполагающих однозначный ответ — «да» или «нет». Последовательно дав ответы на все эти вопросы, мы приходим к правильному выбору. Методологические преимущества дерева принятия решений – в том, что оно структурирует и систематизирует проблему, а итоговое решение принимается на основе логических выводов.

**Наивная байесовская классификация**

Наивные байесовские классификаторы относятся к семейству простых вероятностных классификаторов и берут начало из теоремы Байеса, которая применительно к данному случаю рассматривает функции как независимые (это называется строгим, или наивным, предположением). На практике используется в следующих областях машинного обучения:

* определение спама, приходящего на электронную почту;
* автоматическая привязка новостных статей к тематическим рубрикам;
* выявление эмоциональной окраски текста;
* распознавание лиц и других паттернов на изображениях.

**Метод наименьших квадратов**

Всем, кто хоть немного изучал статистику, знакомо понятие линейной регрессии. К вариантам её реализации относятся и наименьшие квадраты. Обычно с помощью линейной регрессии решают задачи по подгонке прямой, которая проходит через множество точек. Вот как это делается с помощью метода наименьших квадратов: провести прямую, измерить расстояние от неё до каждой из точек (точки и линию соединяют вертикальными отрезками), получившуюся сумму перенести наверх. В результате та кривая, в которой сумма расстояний будет наименьшей, и есть искомая (эта линия пройдёт через точки с нормально распределённым отклонение мотистинного значения).

Линейная функция обычно используется при подборе данных для машинного обучения, а метод наименьших квадратов – для сведения к минимуму погрешностей путем создания метрики ошибок.  
 **Логистическая регрессия**

Логистическая регрессия – это способ определения зависимости между переменными, одна из которых категориально зависима, а другие независимы. Для этого применяется логистическая функция (аккумулятивное логистическое распределение). Практическое значение логистической регрессии заключается в том, что она является мощным статистическим методом предсказания событий, который включает в себя одну или несколько независимых переменных.

Это востребовано в следующих ситуациях:

* кредитный скоринг;
* замеры успешности проводимых рекламных кампаний;
* прогноз прибыли с определённого товара;
* оценкавероятностиземлетрясениявконкретнуюдату.

**Метод опорных векторов (SVM)**

Это целый набор алгоритмов, необходимых для решения задач на классификацию и регрессионный анализ. Исходя из того что объект, находящийся в N-мерном пространстве, относится к одному из двух классов, метод опорных векторов строит гиперплоскость с мерностью (N – 1), чтобы все объекты оказались в одной из двух групп. На бумаге это можно изобразить так: есть точки двух разных видов, и их можно линейно разделить. Кроме сепарации точек, данный метод генерирует гиперплоскость таким образом, чтобы она была максимально удалена от самой близкой точки каждой группы.  
SVM и его модификации помогают решать такие сложные задачи машинного обучения, как сплайсинг ДНК, определение пола человека по фотографии, вывод рекламных баннеров на сайты.

**Метод ансамблей**

Он базируетсяна алгоритмах машинного обучения, генерирующих множество классификаторов и разделяющих все объекты из вновь поступающих данных на основе их усреднения или итогов голосования. Изначально метод ансамблей был частным случаем байесовского усреднения, но затем усложнился и оброс дополнительными алгоритмами:

* бустинг (boosting) – преобразует слабые модели в сильные посредством формирования ансамбля классификаторов (с математической точки зрения это является улучшающим пересечением);
* бэггинг (bagging) – собирает усложнённые классификаторы, при этом параллельно обучая базовые (улучшающее объединение);
* корректирование ошибок выходного кодирования.

Метод ансамблей – более мощный инструмент по сравнению с отдельно стоящими моделями прогнозирования, поскольку:

он сводит к минимуму влияние случайностей, усредняя ошибки каждого базового классификатора;

уменьшает дисперсию, поскольку несколько разных моделей, исходящих из разных гипотез, имеют больше шансов прийти к правильному результату, чем одна отдельно взятая;

исключает выход за рамки множества: если агрегированная гипотеза оказывается вне множества базовых гипотез, то на этапе формирования комбинированной гипотезы оно расширяется при помощи того или иного способа, и гипотеза уже входит в него.

**Алгоритмы кластеризации**

Кластеризация заключается в распределении множества объектов по категориям так, чтобы в каждой категории – кластере –оказались наиболее схожие между собой элементы.

Кластеризировать объекты можно по разным алгоритмам. Чаще всего используютследующие:

* на основе центра тяжести треугольника;
* на базе подключения;
* сокращения размерности;
* плотности (основанные на пространственной кластеризации);
* вероятностные;

Алгоритмы кластеризации используются в биологии (исследование взаимодействия генов в геноме, насчитывающем до нескольких тысяч элементов), социологии (обработка результатов социологических исследований методом Уорда, на выходе дающим кластеры с минимальной дисперсией и примерно одинакового размера) и информационных технологиях.

**Метод главных компонент (PCA)**

Метод главных компонент, или PCA, представляет собой статистическую операцию по ортогональному преобразованию, которая имеет своей целью перевод наблюдений за переменными, которые могут быть как-то взаимосвязаны между собой, в набор главных компонент – значений, которые линейно не коррелированы.

Практические задачи, в которых применяется PCA, – визуализация и большинство процедур сжатия, упрощения, минимизации данных для того, чтобы облегчить процесс обучения. Однако метод главных компонент не годится для ситуаций, когда исходные данные слабо упорядочены (то есть все компоненты метода характеризуются высокой дисперсией). Так что его применимость определяется тем, насколько хорошо изучена и описана предметная область.

**Сингулярное разложение**

В линейной алгебре сингулярное разложение, или SVD, определяется как разложение прямоугольной матрицы, состоящей из комплексных или вещественных чисел. Так, матрицу M размерностью [m\*n] можно разложить таким образом, что M = UΣV, где U и V будут унитарными матрицами, а Σ – диагональной.

Одним из частных случаев сингулярного разложения является метод главных компонент. Самые первые технологии компьютерного зрения разрабатывались на основе SVD и PCA и работали следующим образом: вначале лица (или другие паттерны, которые предстояло найти) представляли в виде суммы базисных компонент, затем уменьшали их размерность, после чего производили их сопоставление с изображениями из выборки. Современные алгоритмы сингулярного разложения в машинном обучении, конечно, значительно сложнее и изощрённее, чем их предшественники, но сутьихвцеломнемизменилась.

**Анализ независимых компонент (ICA)**

Это один из статистических методов, который выявляет скрытые факторы, оказывающие влияние на случайные величины, сигналы и пр. ICA формирует порождающую модель для баз многофакторных данных. Переменные в модели содержат некоторые скрытые переменные, причем нет никакой информации о правилах их смешивания. Эти скрытые переменные являются независимыми компонентами выборки и считаются негауссовскими сигналами.В отличие от анализа главных компонент, который связан с данным методом, анализ независимых компонент более эффективен, особенно в тех случаях, когда классические подходы оказываются бессильны. Он обнаруживает скрытые причины явлений и благодаря этому нашёл широкое применение в самых различных областях – от астрономии и медицины до распознавания речи, автоматического тестирования и анализа динамики финансовых показателей.

**1.3. Разведочный анализ данных**

Разведочный анализ (Exploratory Data Analysis) — это процесс анализа данных, способствующий выдвижению гипотез, когда отсутсвтует представление о связи между переменными или оно недостаточно.

Важно отметить, что РА— это не набор готовых методов анализа, а философия того, как следует анализировать данные, что искать и как интерпретировать.

В разведочном анализе часто используется статистическая графика — графическая техника, используемая для визуализации количественных данных.

Результаты разведочного анализа не используются для выработки управленческих решений. Их назначение — помощь в разработке наилучшей стратегии углубленного анализа, выдвижение гипотез, уточнение особенностей применения тех или иных математических методов и моделей. Без разведочного анализа углубленный анализ данных будет производиться практически «вслепую».

Это необходимо для того, чтобы максимизировать понимание набора данных;

* выявить базовую структуру;
* извлечь важные переменные;
* выявить выбросы и аномалии;
* проверить исходные предположения;
* разработать скупые модели;
* определять оптимальные параметры.

Причина сильной зависимости в графике заключается в том, что основная роль РА — открывать глаза, а графика дает аналитикам возможность заставлять данные раскрывать свои структурные секреты и быть всегда готовыми получить новое, часто непредвиденное, понимание данных. В сочетании с естественными возможностями распознавания образов, которыми все мы обладаем, графика открывает большие шансы для этого.

Довольно распространенные методы использования графики в РА это:

* предоставление необработанных данных с помощью гистограмм, диаграмм рассеяния, вероятностных графиков;
* вычисление простых статистических данных: добавление линии среднего значения и стандартного отклонения; сопоставление нескольких графиков.

К основным методам разведочного анализа относится процедура анализа распределений переменных, [корреляционный анализ](https://wiki.loginom.ru/articles/correlation-analysis.html) c целью поиска [коэффициентов](https://wiki.loginom.ru/articles/correlation-coefficient.html), превосходящих по величине определенные пороговые значения, [факторный анализ](https://wiki.loginom.ru/articles/factorial-analysis.html), [дискриминантный анализ](https://wiki.loginom.ru/articles/linear-discriminant-analysis.html), многомерное шкалирование, визуальный анализ .

Предварительное исследование данных может служить лишь первым этапом в процессе их анализа, и пока результаты не подтверждены на других [выборках](https://wiki.loginom.ru/articles/sample.html) или на независимом множестве данных, их следует воспринимать самое большее как гипотезу. Если результаты разведочного анализа говорят в пользу некоторой модели, то ее правильность можно затем проверить, применив к новым данным.

Корреляционный анализ (Correlation analysis)

Совокупность методов обработки данных с целью обнаружения статистической взаимосвязи между случайными величинами или [признаками](https://wiki.loginom.ru/articles/attribute.html).

Корреляционный анализ для двух случайных величин заключает в себе:

Построение корреляционного поля и составление корреляционной таблицы;

Вычисление выборочных [коэффициентов корреляции](https://wiki.loginom.ru/articles/correlation-coefficient.html) и корреляционных отношений;

Проверка статистической [гипотезы](https://wiki.loginom.ru/articles/hypothesis.html) значимости связи.

Методами корреляционного анализа решаются следующие задачи:

Взаимосвязь. Установление наличия зависимости между двумя признаками и определение ее силы.

Прогнозирование. Предсказание поведения одного признака на основе изменения другого, коррелирующего с первым.

Отбор переменных. Корреляционный анализ позволяет производить выбор набора входных переменных для аналитической модели, в наименьшей степени коррелирующих между собой и в наибольшей степени коррелирующих с [выходной переменной](https://wiki.loginom.ru/articles/output-variable.html). Это позволяет сделать работу аналитических моделей более точной и устойчивой.

Сила корреляционной связи между двумя переменными характеризуется с помощью коэффициента корреляции.

Конкретный вид (линейная, экспоненциальная и т.д.) корреляционной связи и ее параметры устанавливается в рамках [регрессионного анализа](https://wiki.loginom.ru/articles/regression-analysis.html).

Впервые элементы корреляционного анализа в научных исследованиях начал применять французский палеонтолог [Жорж Кювье](https://ru.wikipedia.org/wiki/%D0%9A%D1%8E%D0%B2%D1%8C%D0%B5,_%D0%96%D0%BE%D1%80%D0%B6_%D0%9B%D0%B5%D0%BE%D0%BF%D0%BE%D0%BB%D1%8C%D0%B4), который и ввел в научный обиход термин «корреляция». Значительный вклад в развитие теории корреляционного анализ внесли [Френсис Гальтон](https://ru.wikipedia.org/wiki/%D0%93%D0%B0%D0%BB%D1%8C%D1%82%D0%BE%D0%BD,_%D0%A4%D1%80%D1%8D%D0%BD%D1%81%D0%B8%D1%81), [Карл Пирсон](https://ru.wikipedia.org/wiki/%D0%9F%D0%B8%D1%80%D1%81%D0%BE%D0%BD,_%D0%9A%D0%B0%D1%80%D0%BB), [Чарльз Спирмен](https://ru.wikipedia.org/wiki/%D0%A1%D0%BF%D0%B8%D1%80%D0%BC%D0%B5%D0%BD,_%D0%A7%D0%B0%D1%80%D0%BB%D1%8C%D0%B7_%D0%AD%D0%B4%D0%B2%D0%B0%D1%80%D0%B4), [Морис Кендалл](https://ru.wikipedia.org/wiki/%D0%9A%D0%B5%D0%BD%D0%B4%D0%B0%D0%BB%D0%BB,_%D0%9C%D0%BE%D1%80%D0%B8%D1%81_%D0%94%D0%B6%D0%BE%D1%80%D0%B4%D0%B6) и другие.

**Факторный анализ (Factorial analysis)**

Направление [математической статистики](https://wiki.loginom.ru/articles/mathematical-statistics.html) (один из разделов многомерного статистического анализа), объединяющее вычислительные методы, которые в ряде случаев позволяют получить компактное описание исследуемых явлений на основе обработки больших массивов информации.

От других средств подобного «сжатия информации» (например, распространенных методов статистической группировки объектов) факторный анализ отличается тем, что не опирается на заранее заданный, априорный перечень факторов, влияющих на исследуемые процессы или объекты, а, наоборот, при соблюдении определенных правил и предосторожностей помогает обнаружить наиболее важные из них, причем скрытые (латентные).

Например, аналитик непосредственно наблюдает за множеством различных показателей деятельности предприятий, чтобы выявить закономерности, влияющие на рост производительности труда (уровень квалификации персонала, коэффициент сменности оборудования, электровооруженность труда, «возраст» оборудования, количество мест в столовых и т.д.).

Так или иначе, все факторы, описываемые этими показателями, воздействуют на изучаемый — производительность труда. При этом многие из них связаны между собой, отражая с разных сторон по существу одни и те же явления.

С помощью приемов функционального анализа этих связей ([корреляций](https://wiki.loginom.ru/articles/correlation.html)) удается обнаружить, что на самом деле решающее влияние на рост производительности труда оказывают лишь несколько обобщенных [факторов](https://wiki.loginom.ru/articles/factor.html) (например, размер предприятия, уровень организации труда, характер продукции), непосредственно не наблюдавшихся при исследовании. Собственно, это их действие и проявляется в учитываемых показателях.

Задача состоит в том, чтобы выявить скрытые обобщенные факторы, которые в достаточной для данного исследования степени объясняют изменения изучаемого показателя.

Выявленные факторы позволяют строить аналитические модели с относительно небольшим числом [независимых переменных](https://wiki.loginom.ru/articles/input-variable.html), что упрощает их реализацию и интерпретацию пользователем, снижает вычислительные затраты и время, требуемое на получение решений, а следовательно, повышает оперативность принятия решений на основе результатов анализа.

Основные идеи факторного анализа были заложены английским психологом и антропологом [Фрэнсисом Гальтоном](https://ru.wikipedia.org/wiki/%D0%93%D0%B0%D0%BB%D1%8C%D1%82%D0%BE%D0%BD,_%D0%A4%D1%80%D1%8D%D0%BD%D1%81%D0%B8%D1%81" \t "_blank) во второй половине 19 в.

Дискриминантный линейный анализ (Linear discriminant analysis)

Линейный дискриминантный анализ (ЛДА), а также связанный с ним линейный дискриминант Фишера — методы статистики и [машинного обучения](https://wiki.loginom.ru/articles/machine-learning.html) для нахождения линейных комбинаций признаков, наилучшим образом разделяющих два или более класса объектов или событий. Полученная комбинация может быть использована в качестве линейного классификатора или для сокращения размерности пространства признаков перед последующей [классификацией](https://wiki.loginom.ru/articles/classification.html).

ЛДА представляет собой раздел многомерного статистического анализа, содержанием которого является разработка методов решения задач различения (дискриминации) объектов наблюдения по набору [признаков](https://wiki.loginom.ru/articles/attribute.html). Иными словами, он позволяет изучать различия между двумя и более группами объектов по нескольким признакам одновременно.

ЛДА тесно связан с [дисперсионным анализом](https://wiki.loginom.ru/articles/anova.html) и [регрессионным анализом](https://wiki.loginom.ru/articles/regression-analysis.html), также пытающимися выразить какую-либо зависимую переменную через линейную комбинацию других признаков или измерений. В этих двух методах [зависимая переменная](https://wiki.loginom.ru/articles/output-variable.html) — численная величина, а в ЛДА она является величиной номинальной ([меткой класса](https://wiki.loginom.ru/articles/class-label.html)). Помимо того, ЛДА имеет схожие черты с [методом главных компонент](https://wiki.loginom.ru/articles/principal-component-analysis.html) и [факторным анализом](https://wiki.loginom.ru/articles/factorial-analysis.html), которые ищут линейные комбинации величин, наилучшим образом описывающие данные.

Можно выделить три вида задач дискриминантного анализа:

* определение дискриминирующих признаков (т.е. признаков, которые позволяют отнести наблюдение к той или иной группе);
* построение дискриминирующей функции;
* [прогнозирование](https://wiki.loginom.ru/articles/forecasting.html) будущих событий, связанных с попаданием объекта в ту или иную группу на основе значений его признака (например, предсказание выживаемости пациента после операции).

Основной целью дискриминации является поиск линейной комбинации признаков (называемых дискриминантными признаками), которые позволили бы наилучшим образом разделить рассматриваемые группы.

Рассмотрим линейную функцию, называемую канонической дискриминантной функцией (КДФ):

dkm=a0+a1x1km+a2x2km+...+apxpkm,

где dkm — значение дискриминирующей функции для m-го наблюдения k-й группы, m=1..n, k=1..g, x2ikm — значение дискриминантного признака для m-го наблюдения k-й группы, p — число дискриминантных признаков (размерность многомерного пространства).

С геометрической точки зрения КДФ определяет гиперповерхности в p-мерном пространстве. При p=2 она будет прямой, а при p=3 — плоскостью. Коэффициенты ai первой КДФ выбираются так, чтобы центроиды различных групп как можно больше отличались друг от друга.

Для случая, представленного на рисунке, прямая должна разбить пространство признаков (x1,x2) таким образом, чтобы расстояние между центроидами результирующих подмножеств было максимально возможным.

Коэффициенты второй КДФ выбираются так же, но при этом налагается дополнительное ограничение, чтобы значения второй функции не коррелировали со значениями первой. Аналогично определяются и другие функции. Отсюда следует, что любая КДФ d имеет нулевую внутригрупповую корреляцию с d1,...dg−1.

Если число групп равно g, то число КДФ будет на единицу меньше числа групп. Однако удобно использовать одну, две или три КДФ, поскольку графическое изображение объектов в этом случае будет представлено в одно-, двух- и трехмерных пространствах. Такое представление особенно полезно в случае, когда число дискриминантных признаков велико по сравнению с числом групп.

ЛДА широко используется для решения задач классификации и распознавания образов, понижения размерности входных данных. Хотя он и работает с информацией, которая определяет принадлежность объекта к одному из классов, но сам по себе классификатором не является, а используется как часть линейной классификационной модели.

Преимущество метода — сравнительная простота реализации и интерпретации результатов. Недостаток — чувствительность к распределению исходных данных, когда даже небольшое их изменение приводит к значительным изменениям результатов классификации.

Основные идеи дискриминантного анализа были сформулированы [Роналдом Фишером](https://ru.wikipedia.org/wiki/%D0%A4%D0%B8%D1%88%D0%B5%D1%80,_%D0%A0%D0%BE%D0%BD%D0%B0%D0%BB%D0%B4" \t "_blank) в 1936 г.

# **2. Практическая часть**

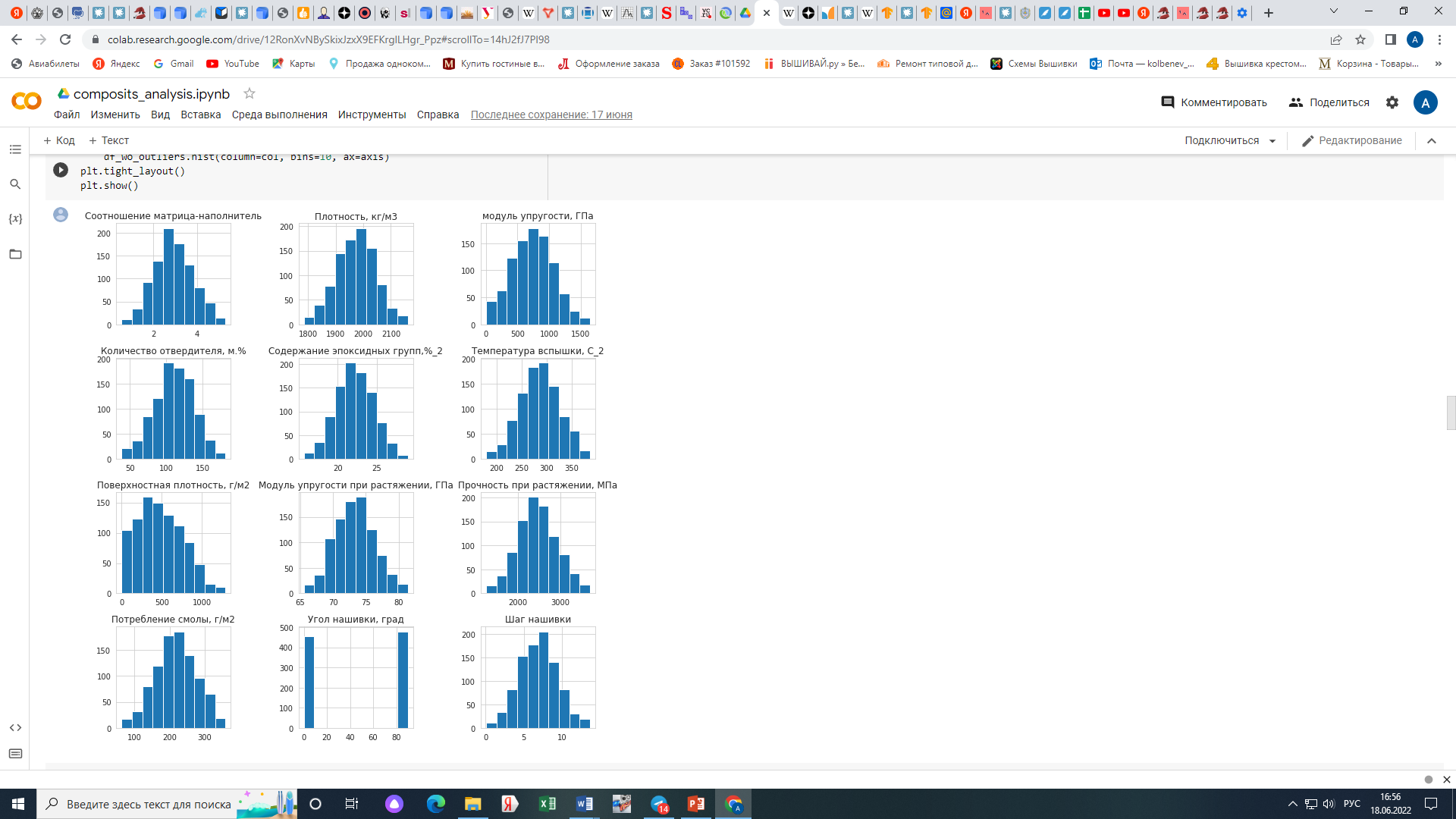
**2.1 Предобработка данных**

Приступим к выполнению задачи по прогнозированию конечных свойств новых композиционных материалов

**На входе** имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

На рисунке №1 представлены гистограммы распределения признаков.

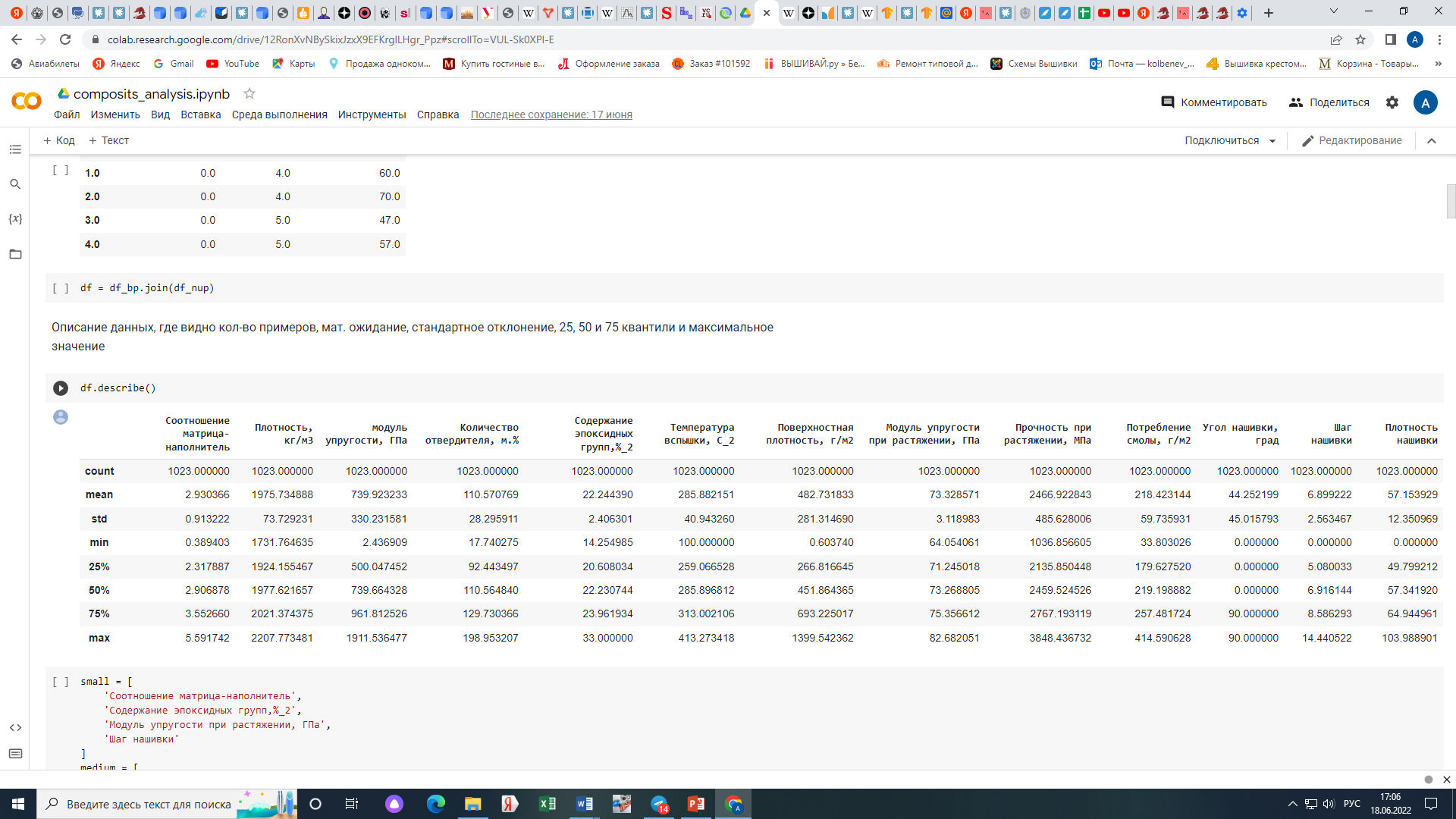
Рисунок №1



По представленным гистограммам видно, что параметры имеют весьма большой разброс.

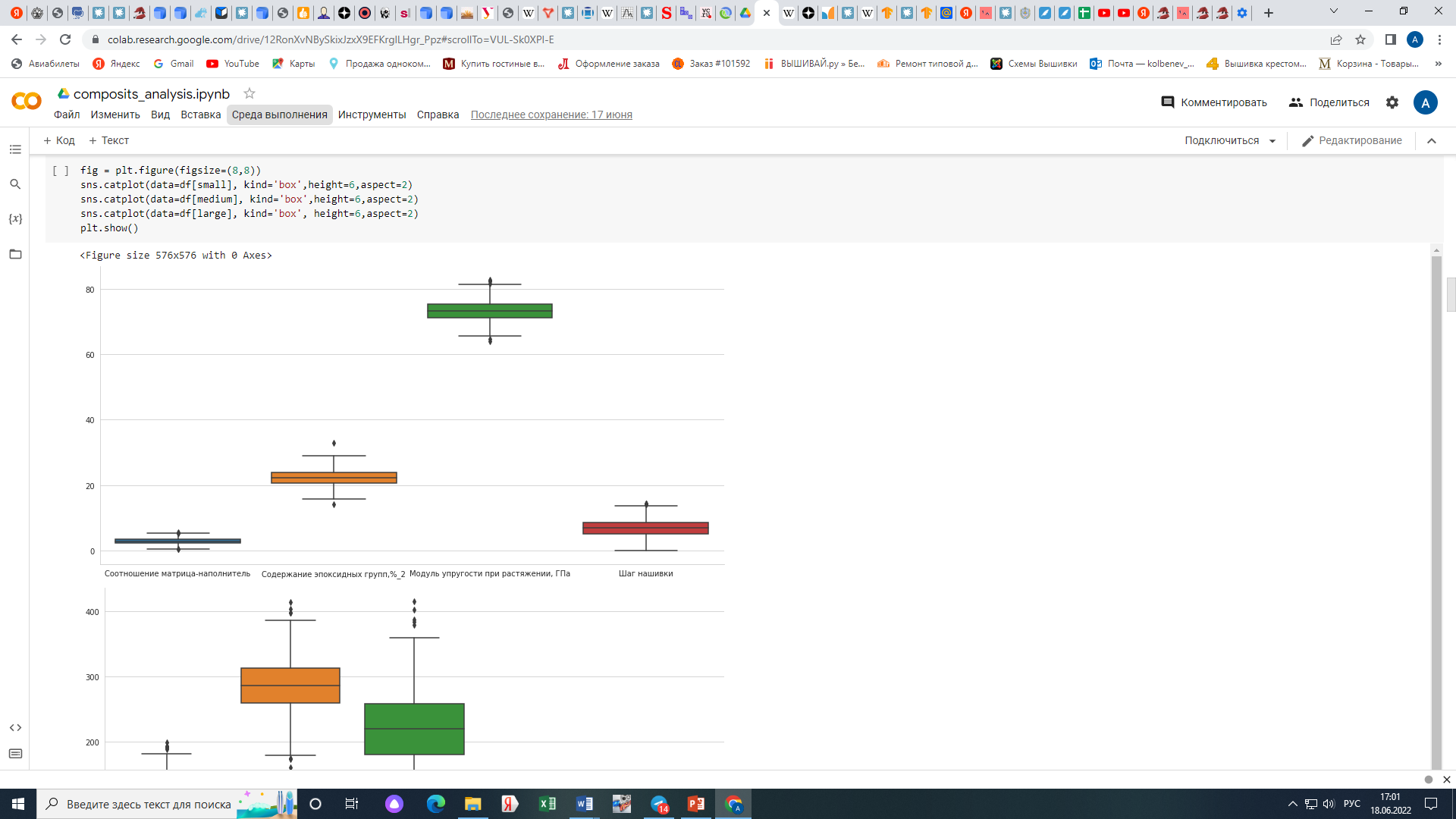
На следующем рисунке представлены максимальные и минимальные значения признаков.

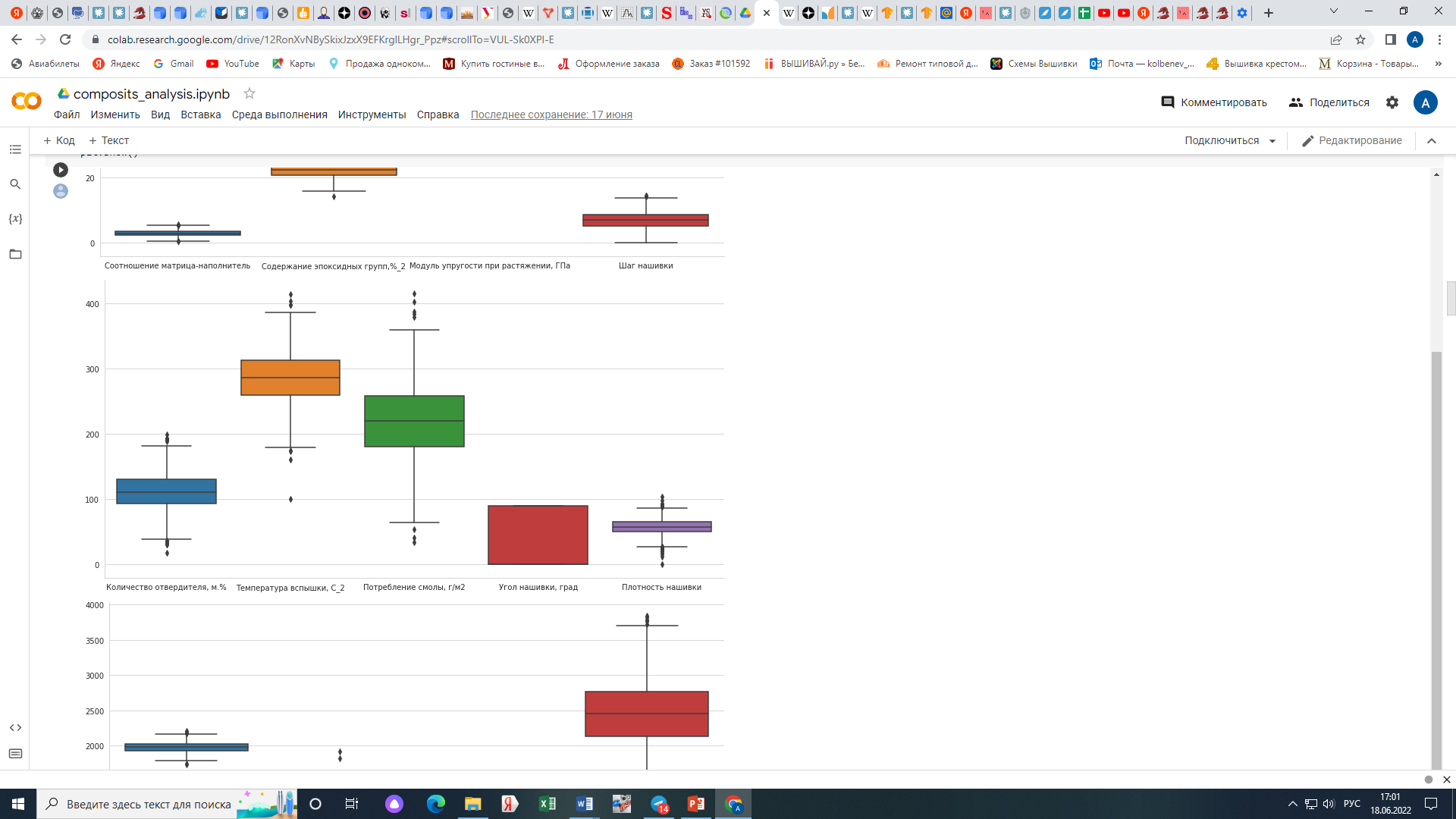
Рисунок №2

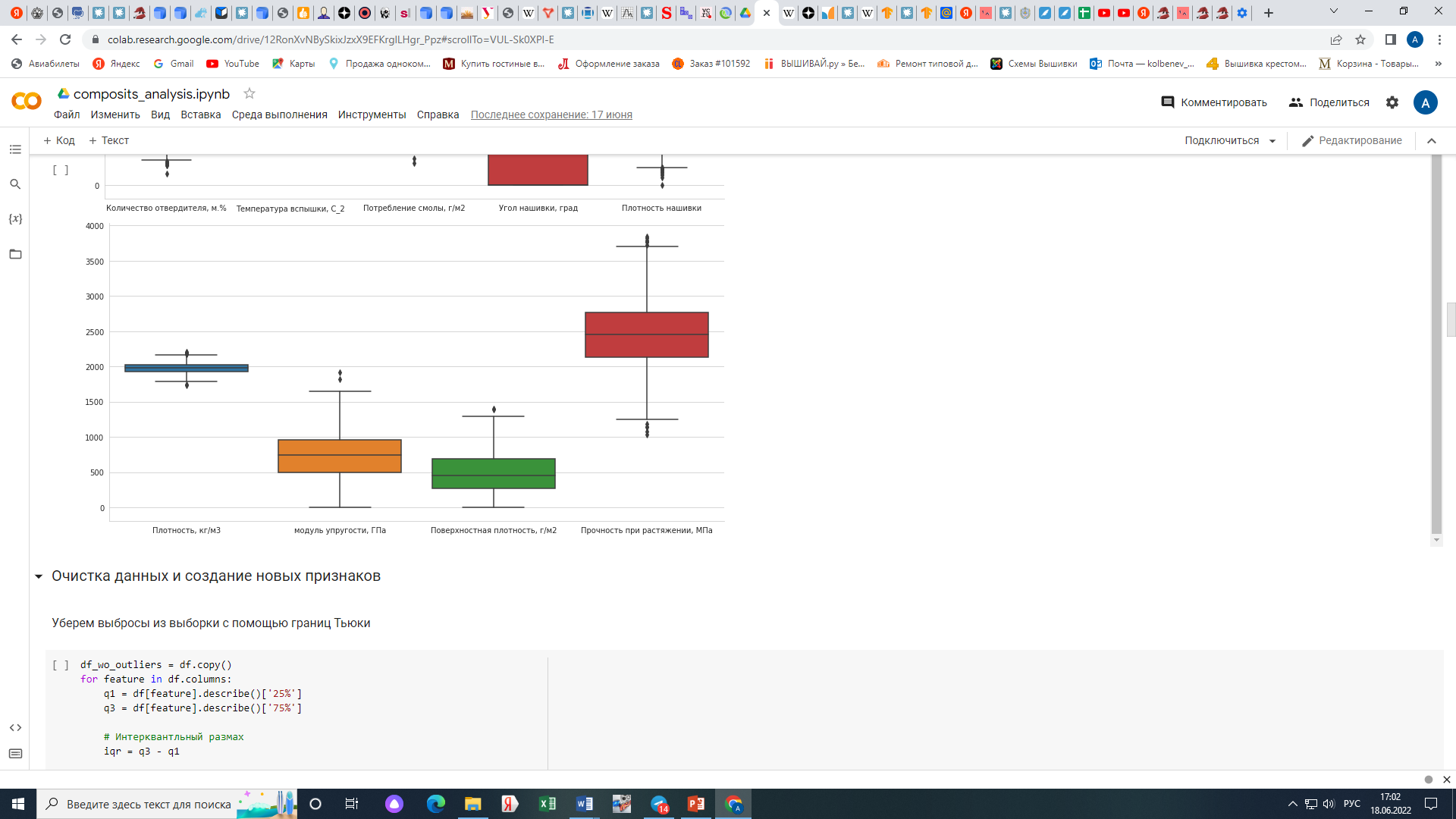


Описанных данных, видно количество примеров, мат. ожидание, стандартное отклонение, 25, 50 и 75 квантили и максимальное значение.

Визуализируем выбросы для различных признаков с помощью диаграмм размаха. Она показывает медиану, 25 и 75 квантили, верхнюю и нижнюю границы, а также выбросы (точки вне прямоугольника и "усов").







Выведем коэффициенты корреляции целевых признаков со всеми признаками, а также корреляционную матрицу для признаков.

Количество отвердителя, м.% -0.048208

Плотность, кг/м3 -0.021316

Соотношение матрица-наполнитель -0.019248

Шаг нашивки -0.007546

Прочность при растяжении, МПа -0.001422

Плотность нашивки 0.013445

Температура вспышки, С\_2 0.015361

модуль упругости, ГПа 0.021086

Поверхностная плотность, г/м2 0.029643

Угол нашивки, град 0.034193

Потребление смолы, г/м2 0.055023

Содержание эпоксидных групп,%\_2 0.055271

Модуль упругости при растяжении, ГПа 1.000000

Name: Модуль упругости при растяжении, ГПа, dtype: float64

Плотность, кг/м3 -0.076305

Количество отвердителя, м.% -0.063256

Шаг нашивки -0.061219

Поверхностная плотность, г/м2 -0.031777

Содержание эпоксидных групп,%\_2 -0.007097

Температура вспышки, С\_2 -0.004877

Модуль упругости при растяжении, ГПа -0.001422

Плотность нашивки 0.016894

Угол нашивки, град 0.017669

Соотношение матрица-наполнитель 0.024488

Потребление смолы, г/м2 0.027949

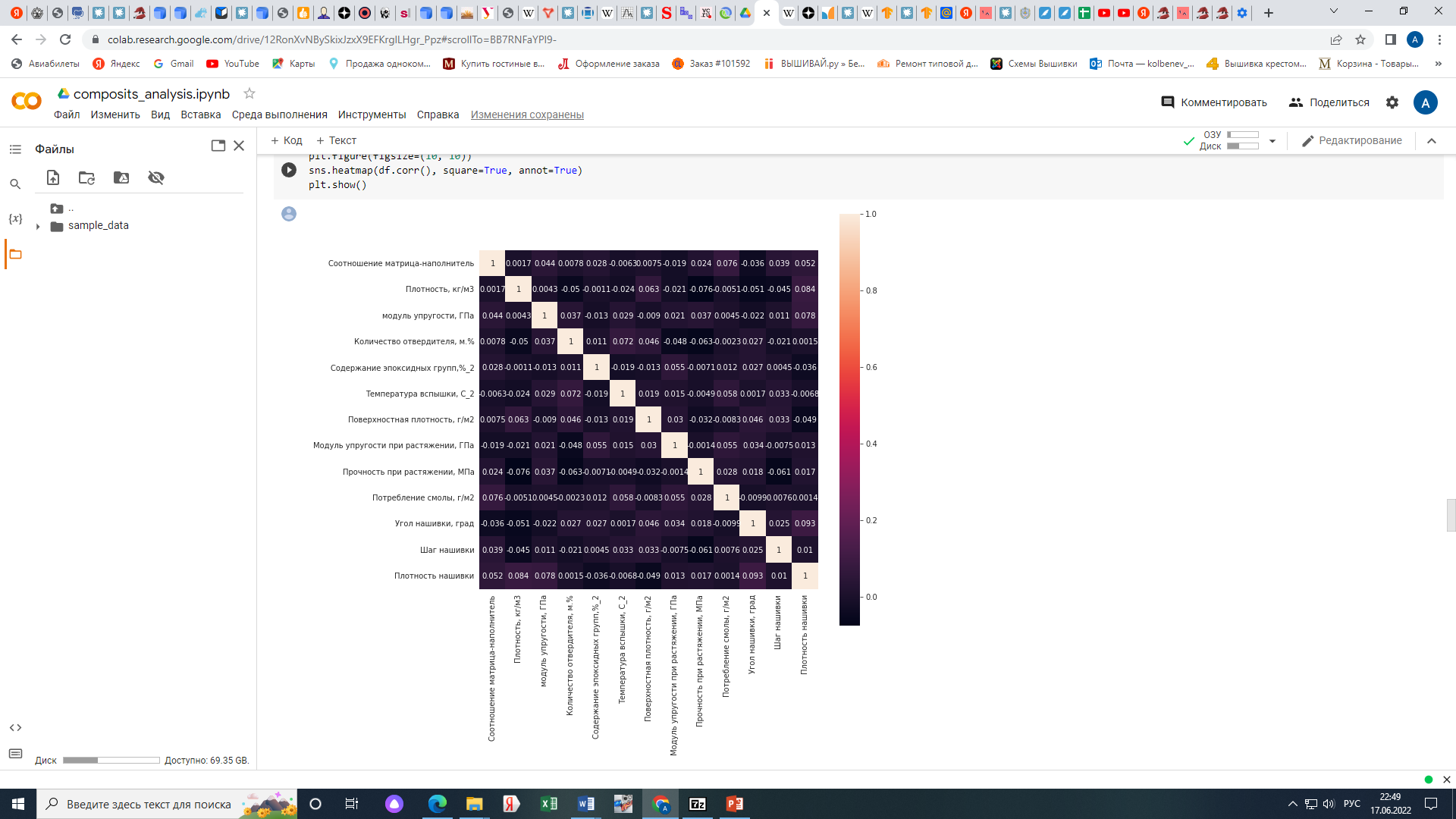
модуль упругости, ГПа 0.036641

Прочность при растяжении, МПа 1.000000

Name: Прочность при растяжении, МПа, dtype: float64

Из показанных значений видно, что у целевых признаков нет сильной корреляции ни с одним из исходных признаков.

Наиболее наглядно это можно увидеть на следующем графике



**2.2 Разработка и обучение модели**

* При разработке и обучении модели использовались следующие методы: Метод опорных векторов (Support Vector Machines — SVM)
* Линейная регрессия
* RandomForestRegressor

Из выбранных моделей наилучший результат показали модели: Random Forest, SVM и XGBoost.

Model: LinearRegression, prediction time: 0.0001, MSE: 1.013406860858747

Model: SVM, prediction time: 0.0066, MSE: 1.1056956872301062

Model: RandomForest, prediction time: 0.0087, MSE: 1.0116117089837213

Model: XGBoost, prediction time: 0.0006, MSE: 1.227911619608587

В результате обучения нейронная сеть показала ошибку MSE 1.066 на тестовом уровне.

# Заключение

При решении данной задачи не удалось достигнуть желаемых результатов, так как признаки плохо между собой коррелируют ни один из признаков не приближается близко к параметрам упругости и прочности. Соответственно не удалось подобрать оптимальную комбинацию признаков. Поэтому модель плохо предсказывает свойства материала.

# Список литературы

1. Композиционные материалы волокнистого строения.К., 1980.
2. Конкин А.А., Углеродные и другие жаростойкие волокнистые материалы, М., 1974.
3. Композиционные материалы, пер. с англ., т.1-8, М., 1998.
4. Наполнители для полимерных композиционных материалов, пер. с англ., М., 2005.

