

## **Выпускная квалификационная работа по курсу «Data Science PRO 2.0»**

Слушатель: Тимошевский Александр Станиславович

**Тема: "Прогнозирование конечных свойств новых материалов (композиционных материалов)".**

### **Задание.**

1. Обучить алгоритм машинного обучения, который будет определять значения:
  - Модуль упругости при растяжении, ГПа;
  - Прочность при растяжении, МПа.
1. Написать нейронную сеть, которая будет рекомендовать:
  - Соотношение матрица-наполнитель.
1. Написать приложение, которое будет выдавать прогноз, полученный в задании 2.

### **Описание:**

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 растягивается. 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.

### **Актуальность:**

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

### **Часть 1. Описательный и разведочный анализ, предоставленных для исследования данных**

Данные для исследования (датасет со свойствами композитов) размещены в сети Интернет по адресу: [https://drive.google.com/file/d/1B1s5gBlvgU81H9GGolLQVw\\_SOi-vyNf2/view?usp=sharing](https://drive.google.com/file/d/1B1s5gBlvgU81H9GGolLQVw_SOi-vyNf2/view?usp=sharing).

Данные разделены на два файла:

- файл "X\_bp.xlsx";
- файл "X\_nipr.xlsx". Формат ".xlsx" это формат файла, используемый для хранения электронных таблиц в программе Microsoft Excel.

Загружаем библиотеки необходимые для проведения исследования

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import missingno as msno

from sklearn.preprocessing import OneHotEncoder, MinMaxScaler
import warnings
warnings.filterwarnings('ignore')
```

С помощью библиотеки pandas осуществляем чтение данных из файлов (датасетов)

```
data_frame1 = pd.read_excel("X_bp.xlsx")
data_frame2 = pd.read_excel("X_nipr.xlsx")
```

Данные загружены и присвоены переменным data\_frame1 и data\_frame2.

### Рассмотрим общие сведения о data\_frame1

Выведим 3 первых строки с данными из data\_frame1

```
data_frame1.head(3)

    Уннамед: 0 Соотношение матрица-наполнитель Плотность, кг/м3 \
0          0           1.857143        2030.0
1          1           1.857143        2030.0
2          2           1.857143        2030.0

    модуль упругости, ГПа Количество отвердителя, м.% \
0      738.736842        30.0
1      738.736842        50.0
2      738.736842        49.9

    Содержание эпоксидных групп,%_2 Температура вспышки, С_2 \
0            22.267857        100.000000
1            23.750000        284.615385
2            33.000000        284.615385

    Поверхностная плотность, г/м2 Модуль упругости при растяжении, ГПа \
0            210.0             70.0
```

1	210.0	70.0
2	210.0	70.0
0	Прочность при растяжении, МПа	Потребление смолы, г/м2
1	3000.0	220.0
2	3000.0	220.0
	3000.0	220.0

Выведим 3 последних строки с данными из data\_frame1

	data_frame1.tail(3)
1020	Unnamed: 0 Соотношение матрица-наполнитель Плотность, кг/м3 \
1021	1020 3.280604 1972.372865
1022	1021 3.705351 2066.799773
	1022 3.808020 1890.413468
1020	модуль упругости, ГПа Количество отвердителя, м.% \
1021	416.836524 110.533477
1022	741.475517 141.397963
	417.316232 129.183416
1020	Содержание эпоксидных групп,%_2 Температура вспышки, С_2 \
1021	23.957502 248.423047
1022	19.246945 275.779840
	27.474763 300.952708
1020	Поверхностная плотность, г/м2 Модуль упругости при растяжении, ГПа \
74.734344	740.142791
1021	641.468152
74.042708	
1022	758.747882
74.309704	
1020	Прочность при растяжении, МПа Потребление смолы, г/м2
1021	2662.906040 236.606764
1022	2071.715856 197.126067
	2856.328932 194.754342

Выведим 3 случайных строки с данными из data\_frame1

	data_frame1.sample(3)
383	Unnamed: 0 Соотношение матрица-наполнитель Плотность, кг/м3 \
	383 2.709714 1963.169297

971	971	2.169074	1889.513179
618	618	3.276517	1911.245306
383	модуль упругости, ГПа	Количество отвердителя, м.%	\
971	766.471349	135.144779	
618	803.346469	132.528295	
383	213.466388	78.847811	
383	Содержание эпоксидных групп,%_2	Температура вспышки, С_2	\
971	24.693930	222.276036	
618	24.075555	323.153659	
383	21.778934	226.990371	
383	Поверхностная плотность, г/м2	Модуль упругости при растяжении,	
73.677834	\ ГПа		
971	376.967176		
75.336917			
618	450.429300		
72.695218	113.590494		
383	Прочность при растяжении, МПа	Потребление смолы, г/м2	
971	1892.581263	269.239867	
618	2147.576085	244.606070	
383	2734.030447	223.770443	

Выводим информацию об общем размере data\_frame1 (количество строк и количество столбцов)

```
data_frame1.shape
(1023, 11)
```

Итого таблица № 1 с предложенными данными содержит 11 столбцов и 1023 строки.

Получаем общее описание data\_frame1 с помощью метода pandas.DataFrame.info()

```
data_frame1.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1023 entries, 0 to 1022
Data columns (total 11 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   Unnamed: 0        1023 non-null   int64  
 1   Соотношение матрица-наполнитель 1023 non-null   float64 
 2   Плотность, кг/м3                 1023 non-null   float64 
 3   модуль упругости, ГПа            1023 non-null   float64 
 4   Количество отвердителя, м.%     1023 non-null   float64
```

```

5 Содержание эпоксидных групп, %_2          1023 non-null   float64
6 Температура вспышки, C_2                  1023 non-null   float64
7 Поверхностная плотность, г/м2            1023 non-null   float64
8 Модуль упругости при растяжении, ГПа    1023 non-null   float64
9 Прочность при растяжении, МПа           1023 non-null   float64
10 Потребление смолы, г/м2                 1023 non-null   float64
dtypes: float64(10), int64(1)
memory usage: 88.0 KB

```

Из общего описания делаем дополнительный вывод о том, что в таблице представлены 2 типа данных:

- int64 - это 32-битный целочисленный тип данных. Он используется для представления целых чисел.
- float32 - это 64-битный формат представления чисел с плавающей точкой. Он используется для хранения вещественных чисел.

Кроме того, отмечаем что столбец "Unnamed:0" содержит нумерацию строк электронной таблицы.

### **Рассмотрим общие сведения о data\_frame2**

Выведим 3 первых строки с данными из data\_frame2

```
data_frame2.head(3)
```

	Unnamed: 0	Угол нашивки, град	Шаг нашивки	Плотность нашивки
0	0	0	4.0	57.0
1	1	0	4.0	60.0
2	2	0	4.0	70.0

Выведим 3 последних строки с данными из data\_frame2

```
data_frame2.tail(3)
```

	Unnamed: 0	Угол нашивки, град	Шаг нашивки	Плотность нашивки
1037	1037	90	9.800926	72.858286
1038	1038	90	10.079859	65.519479
1039	1039	90	9.021043	66.920143

Выведим 3 случайных строки с данными из data\_frame2

```
data_frame2.sample(3)
```

	Unnamed: 0	Угол нашивки, град	Шаг нашивки	Плотность нашивки
50	50	0	6.303773	72.152019
407	407	0	12.634844	37.547373
268	268	0	4.803995	79.931964

Выводим информацию об общем размере data\_frame2 (количество строк и количество столбцов)

```
data_frame2.shape  
(1040, 4)
```

Итого таблица № 2 с предложенными данными содержит 4 столбца и 1040 строк.

Получаем общее описание data\_frame2 с помощью метода pandas.DataFrame.info()

```
data_frame2.info()  
  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 1040 entries, 0 to 1039  
Data columns (total 4 columns):  
 #   Column           Non-Null Count  Dtype     
---  --    
 0   Unnamed: 0        1040 non-null    int64    
 1   Угол нашивки, град 1040 non-null    int64    
 2   Шаг нашивки      1040 non-null    float64  
 3   Плотность нашивки 1040 non-null    float64  
dtypes: float64(2), int64(2)  
memory usage: 32.6 KB
```

Из общего описания делаем дополнительный вывод о том, что в таблице представлены 2 типа данных:

int64 - это 32-битный целочисленный тип данных. Он используется для представления целых чисел. float32 - это 64-битный формат представления чисел с плавающей точкой. Он используется для хранения вещественных чисел.

Кроме того, отмечаем что столбец "Unnamed:0" содержит нумерацию строк электронной таблицы.

### Объединение предоставленных данных

В соответствии с требованиями к ВКР, предоставленные данные должны быть объединены в одну таблицу. Объединение должно производиться по индексам, тип объединения - INNER. Это самый распространенный тип объединения. С его помощью происходит объединение записей из двух таблиц по связующему полю, если оно содержит одинаковые значения в обеих таблицах.

Перед объединением таблицы № 1 (data\_frame1) и таблицы № 2(data\_frame2) исключаем из них столбцы "Unnamed: 0" в которых содержится нумерация строк. Данные столбцы не влияют на результаты исследования.

```
data_frame1.drop(["Unnamed: 0"], axis=1, inplace=True)  
  
# Проверяем, что столбец "Unnamed: 0" успешно удален из таблицы № 1  
data_frame1.tail(5)
```

	Соотношение матрица-наполнитель	Плотность, кг/м3	\
1018		2.271346	1952.087902
1019		3.444022	2050.089171
1020		3.280604	1972.372865
1021		3.705351	2066.799773
1022		3.808020	1890.413468
	модуль упругости, ГПа	Количество отвердителя, м.%	\
1018	912.855545		86.992183
1019	444.732634		145.981978
1020	416.836524		110.533477
1021	741.475517		141.397963
1022	417.316232		129.183416
	Содержание эпоксидных групп,%_2	Температура вспышки, С_2	\
1018	20.123249		324.774576
1019	19.599769		254.215401
1020	23.957502		248.423047
1021	19.246945		275.779840
1022	27.474763		300.952708
	Поверхностная плотность, г/м2	Модуль упругости при растяжении, ГПа	\
1018		209.198700	
73.090961			
1019		350.660830	
72.920827			
1020		740.142791	
74.734344			
1021		641.468152	
74.042708			
1022		758.747882	
74.309704			
	Прочность при растяжении, МПа	Потребление смолы, г/м2	
1018	2387.292495		125.007669
1019	2360.392784		117.730099
1020	2662.906040		236.606764
1021	2071.715856		197.126067
1022	2856.328932		194.754342

# Проверяем, что столбец "Unnamed: 0" успешно удален из таблицы № 1  
data\_frame1.shape

(1023, 10)

Удостоверились, что исключен столбец "Unnamed: 0". Столбцов стало 10, а количество строк прежнее - 1023.

```
data_frame2.drop(["Unnamed: 0"],axis=1, inplace=True)
```

```
# Проверяем, что столбец "Unnamed: 0" успешно удален из таблицы № 2  
data_frame2.tail(5)
```

	Угол нашивки, град	Шаг нашивки	Плотность нашивки
1035	90	8.088111	47.759177
1036	90	7.619138	66.931932
1037	90	9.800926	72.858286
1038	90	10.079859	65.519479
1039	90	9.021043	66.920143

```
# Проверяем, что столбец "Unnamed: 0" успешно удален из таблицы № 2  
data_frame2.shape
```

```
(1040, 3)
```

Удостоверились, что исключен столбец "Unnamed: 0". Столбцов стало 3, а количество строк прежнее - 1040.

Осуществляем объединение таблиц

Метод DataFrame.merge() модуля pandas выполняет объединение DataFrame или именованные объекты Series с помощью соединения в стиле базы данных.

```
df = data_frame1.merge(data_frame2, left_index = True, right_index = True, how = 'inner')
```

```
df.head(3)
```

	Соотношение матрица-наполнитель упругости, ГПа	Плотность, кг/м3	модуль
0	1.857143	2030.0	
1	1.857143	2030.0	
2	1.857143	2030.0	
738.736842			

	Количество отвердителя, м.%	Содержание эпоксидных групп,%_2
0	30.0	22.267857
1	50.0	23.750000
2	49.9	33.000000

	Температура вспышки, С_2	Поверхностная плотность, г/м2
0	100.000000	210.0
1	284.615385	210.0
2	284.615385	210.0

	Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа
0	70.0	3000.0

1	70.0	3000.0
2	70.0	3000.0

Потребление смолы, г/м2 Угол нашивки, град Шаг нашивки Плотность нашивки

0	220.0	0	4.0
57.0			
1	220.0	0	4.0
60.0			
2	220.0	0	4.0
70.0			

df.tail(3)

	Соотношение матрица-наполнитель	Плотность, кг/м3	\
1020	3.280604	1972.372865	
1021	3.705351	2066.799773	
1022	3.808020	1890.413468	

	модуль упругости, ГПа	Количество отвердителя, м.%	\
1020	416.836524	110.533477	
1021	741.475517	141.397963	
1022	417.316232	129.183416	

	Содержание эпоксидных групп,%_2	Температура вспышки, С_2	\
1020	23.957502	248.423047	
1021	19.246945	275.779840	
1022	27.474763	300.952708	

	Поверхностная плотность, г/м2	Модуль упругости при растяжении, ГПа	\
1020	740.142791	74.734344	
1021	641.468152	74.042708	
1022	758.747882	74.309704	

	Прочность при растяжении, МПа	Потребление смолы, г/м2	\
1020	2662.906040	236.606764	
1021	2071.715856	197.126067	
1022	2856.328932	194.754342	

	Угол нашивки, град	Шаг нашивки	Плотность нашивки
1020	90	4.161154	67.629684
1021	90	6.313201	58.261074
1022	90	6.078902	77.434468

df.sample(3)

Соотношение матрица-наполнитель упругости, ГПа		Плотность, кг/м3	модуль
924		1.862875	1993.662575
102.490025			
17		4.193548	1950.000000
506.000000			
750		3.273888	1968.408963
928.020607			
Количество отвердителя, м.%		Содержание эпоксидных групп,%_2	\
924	142.797271		24.154225
17	129.000000		21.250000
750	104.943294		22.980391
Температура вспышки, С_2		Поверхностная плотность, г/м2	\
924	354.890074		26.053941
17	300.000000		380.000000
750	291.946090		339.140842
Модуль упругости при растяжении, ГПа		Прочность при растяжении,	
МПа			\
924		74.553726	
2100.602732			
17		75.000000	
1800.000000			
750		74.847458	
2648.655464			
Потребление смолы, г/м2		Угол нашивки, град	Шаг нашивки
924	211.743890	90	6.759424
17	120.000000	0	10.000000
750	282.835721	90	7.083955
Плотность нашивки			
924	60.888620		
17	60.000000		
750	52.843871		

df.shape

(1023, 13)

Удостоверились, что с помощью метода pandas.DataFrame.merge() произведено объединение таблиц № 1 и 2. Объединение произведено по индексам, методом INNER. Итоговая таблица имеет размер 13 столбцов и 1023 строки. Отмечаем, что 17 строк из таблицы № 2 не вошли в объединенную таблицу, так как за основу объединения бралась таблица № 1, в которой отсутствовала информация по этим 17 строкам.

### Описательный анализ объединенных из таблиц данных

```
# Выводим данные датафрейма
```

```
df
```

	Соотношение матрица-наполнитель	Плотность, кг/м3	\
0	1.857143	2030.000000	
1	1.857143	2030.000000	
2	1.857143	2030.000000	
3	1.857143	2030.000000	
4	2.771331	2030.000000	
..	..	..	
1018	2.271346	1952.087902	
1019	3.444022	2050.089171	
1020	3.280604	1972.372865	
1021	3.705351	2066.799773	
1022	3.808020	1890.413468	

	модуль упругости, ГПа	Количество отвердителя, м.%	\
0	738.736842	30.000000	
1	738.736842	50.000000	
2	738.736842	49.900000	
3	738.736842	129.000000	
4	753.000000	111.860000	
..	..	..	
1018	912.855545	86.992183	
1019	444.732634	145.981978	
1020	416.836524	110.533477	
1021	741.475517	141.397963	
1022	417.316232	129.183416	

	Содержание эпоксидных групп,%_2	Температура вспышки, С_2	\
0	22.267857	100.000000	
1	23.750000	284.615385	
2	33.000000	284.615385	
3	21.250000	300.000000	
4	22.267857	284.615385	
..	..	..	
1018	20.123249	324.774576	
1019	19.599769	254.215401	
1020	23.957502	248.423047	
1021	19.246945	275.779840	
1022	27.474763	300.952708	

	Поверхностная плотность, г/м2	Модуль упругости при растяжении, ГПа	\
0		210.000000	
70.000000		210.000000	
1		210.000000	
70.000000		210.000000	
2		210.000000	
70.000000			

3	210.000000
70.000000	
4	210.000000
70.000000	
...	...
...	
1018	209.198700
73.090961	
1019	350.660830
72.920827	
1020	740.142791
74.734344	
1021	641.468152
74.042708	
1022	758.747882
74.309704	

	Прочность при растяжении, МПа	Потребление смолы, г/м <sup>2</sup>	\
0	3000.000000	220.000000	
1	3000.000000	220.000000	
2	3000.000000	220.000000	
3	3000.000000	220.000000	
4	3000.000000	220.000000	
...	...	...	
1018	2387.292495	125.007669	
1019	2360.392784	117.730099	
1020	2662.906040	236.606764	
1021	2071.715856	197.126067	
1022	2856.328932	194.754342	

	Угол нашивки, град	Шаг нашивки	Плотность нашивки
0	0	4.000000	57.000000
1	0	4.000000	60.000000
2	0	4.000000	70.000000
3	0	5.000000	47.000000
4	0	5.000000	57.000000
...	...	...	...
1018	90	9.076380	47.019770
1019	90	10.565614	53.750790
1020	90	4.161154	67.629684
1021	90	6.313201	58.261074
1022	90	6.078902	77.434468

[1023 rows x 13 columns]

Название столбцов и их описание.

Названия столбцов

Описание полей

Соотношение матрица-наполнитель

матрица - это связующее вещество, в которое

Названия столбцов	Описание полей
Плотность, кг/м3	внедрен наполнитель (волокна, частицы, слои). Соотношение может быть выражено в процентах или в виде объемного или весового соотношения.
модуль упругости, ГПа	физическая величина, характеризующая массу вещества, содержащегося в единице объема. Выражен в килограмме на кубический метр.
Количество отвердителя, м.%	физическая величина, характеризующая способность твердого тела сопротивляться упругой деформации приложении к нему силы. Выражен в Гигапаскаль.
Содержание эпоксидных групп, %_2	отвердитель – это химическое вещество, которое добавляется к лакокрасочным материалам, эпоксидным смолам и другим реакционноспособным олигомерам для ускорения или инициирования процесса отверждения (застывания, полимеризации).
Температура вспышки, С_2	эпоксидные группы – это структурные единицы в молекулах эпоксидных смол, которые обеспечивают их способность к отверждению под действием отвердителей.
Поверхностная плотность, г/м2	это самая низкая температура, при которой пары над поверхностью горючего вещества способны вспыхнуть при поднесении к ним источника зажигания, но устойчивое горение после удаления источника зажигания не наблюдается.
Модуль упругости при растяжении, ГПа	величина, характеризующая массу вещества, приходящуюся на единицу площади. Выражен в граммах на метр кубический
Прочность при растяжении, МПа	ценивает упругость жестких или твердых материалов, которая является соотношением между деформацией материала и силой, необходимой для его деформации. Выражен в Гигапаскаль.
Потребление смолы, г/м2	максимальное напряжение, которое материал может выдержать при растягивающей нагрузке до начала разрушения. Это важный показатель для оценки прочности и надежности материалов. Выражен в Мегапаскаль.
Угол нашивки, град	потребление эпоксидной смолы. Выражен в граммах на квадратный метр
Шаг нашивки	угол нашивки. Выражен в градусах
	шаг нашивки

Названия столбцов	Описание полей
Плотность нашивки	плотность нашивки

Вычисляем общее количество значений в таблице.

```
# Общее количество значений равно произведению количества строк на
# столбцы
df.size

13299
```

Выводим общую информацию о таблице

```
df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 1023 entries, 0 to 1022
Data columns (total 13 columns):
 #   Column           Non-Null Count  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
dtypes: float64(12), int64(1)
memory usage: 111.9 KB
```

Метод pd.DataFrame.info() указывает на отсутствует пропущенных значений. В таблице используются числовые типы данных - для целых чисел (int64) и для чисел с плавающей точкой (float64).

Дополнительно исследуем пропущенные таблицы в датасете

```
df.isna().sum()

Соотношение матрица-наполнитель      0
Плотность, кг/м3                      0
модуль упругости, ГПа                 0
Количество отвердителя, м.%           0
Содержание эпоксидных групп,%_2       0
Температура вспышки, С_2              0
```

```
Поверхностная плотность, г/м2          0
Модуль упругости при растяжении, ГПа    0
Прочность при растяжении, МПа           0
Потребление смолы, г/м2                 0
Угол нашивки, град                       0
Шаг нашивки                             0
Плотность нашивки                      0
dtype: int64
```

```
df.isnull().sum()
```

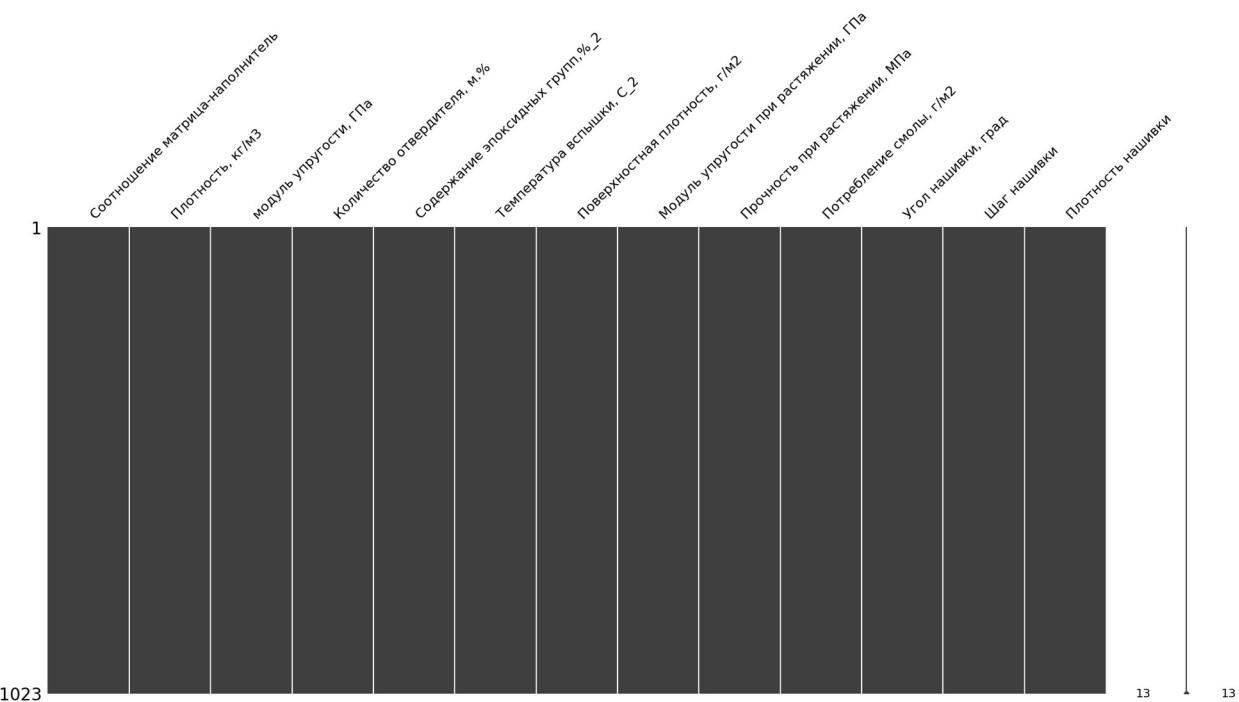
```
Соотношение матрица-наполнитель        0
Плотность, кг/м3                         0
модуль упругости, ГПа                   0
Количество отвердителя, м.%            0
Содержание эпоксидных групп,%_2       0
Температура вспышки, С_2                0
Поверхностная плотность, г/м2          0
Модуль упругости при растяжении, ГПа    0
Прочность при растяжении, МПа           0
Потребление смолы, г/м2                 0
Угол нашивки, град                       0
Шаг нашивки                             0
Плотность нашивки                      0
dtype: int64
```

```
df.isnull().sum().sum()
```

```
np.int64(0)
```

С помощью библиотеки missingno дополнительно исследуем пропуски в датасете

```
msno.matrix(df)
<Axes: >
```



Пропущенные значения в датасете отсутствуют. Совершение каких-либо действий по удалению пропущенных значений не требуется.

Проверяем уникальные значения в таблице

```
df.unique()
```

Соотношение матрица-наполнитель	1014
Плотность, кг/м3	1013
модуль упругости, ГПа	1020
Количество отвердителя, м.%	1005
Содержание эпоксидных групп,%_2	1004
Температура вспышки, С_2	1003
Поверхностная плотность, г/м2	1004
Модуль упругости при растяжении, ГПа	1004
Прочность при растяжении, МПа	1004
Потребление смолы, г/м2	1003
Угол нашивки, град	2
Шаг нашивки	989
Плотность нашивки	988

Отмечаем, что большинство переменных имеют уникальные значения и какая-либо общая тенденция отсутствует. Так же обнаруживаем что столбец "Угол нашивки, град" имеет только 2 уникальных значения в 1023 строках.

```
# Выводим уникальные значения переменной "Угол нашивки, град"
df["Угол нашивки, град"].unique()
```

```
array([ 0, 90])
```

Такими значениями являются - 0 и 90. Как следует из наименования столбца единица измерения градусы.

С помощью метода pd.DataFrame.describe() выводим описательную статистику датасета. Также преобразуем таблицу с помощью транспонирования для более удобной оценки параметров.

```
df.describe().T
```

	count	mean	std
\ Соотношение матрица-наполнитель	1023.0	2.930366	0.913222
Плотность, кг/м3	1023.0	1975.734888	73.729231
модуль упругости, ГПа	1023.0	739.923233	330.231581
Количество отвердителя, м.%	1023.0	110.570769	28.295911
Содержание эпоксидных групп,%_2	1023.0	22.244390	2.406301
Температура вспышки, С_2	1023.0	285.882151	40.943260
Поверхностная плотность, г/м2	1023.0	482.731833	281.314690
Модуль упругости при растяжении, ГПа	1023.0	73.328571	3.118983
Прочность при растяжении, МПа	1023.0	2466.922843	485.628006
Потребление смолы, г/м2	1023.0	218.423144	59.735931
Угол нашивки, град	1023.0	44.252199	45.015793
Шаг нашивки	1023.0	6.899222	2.563467
Плотность нашивки	1023.0	57.153929	12.350969
50% \			
Соотношение матрица-наполнитель	0.389403	2.317887	
Плотность, кг/м3	1731.764635	1924.155467	
модуль упругости, ГПа	2.436909	500.047452	
Количество отвердителя, м.%	17.740275	92.443497	
Содержание эпоксидных групп,%_2	14.254985	20.608034	

22.230744		
Температура вспышки, С_2	100.000000	259.066528
285.896812		
Поверхностная плотность, г/м2	0.603740	266.816645
451.864365		
Модуль упругости при растяжении, ГПа	64.054061	71.245018
73.268805		
Прочность при растяжении, МПа	1036.856605	2135.850448
2459.524526		
Потребление смолы, г/м2	33.803026	179.627520
219.198882		
Угол нашивки, град	0.000000	0.000000
0.000000		
Шаг нашивки	0.000000	5.080033
6.916144		
Плотность нашивки	0.000000	49.799212
57.341920		
	75%	max
Соотношение матрица-наполнитель	3.552660	5.591742
Плотность, кг/м3	2021.374375	2207.773481
модуль упругости, ГПа	961.812526	1911.536477
Количество отвердителя, м.%	129.730366	198.953207
Содержание эпоксидных групп,%_2	23.961934	33.000000
Температура вспышки, С_2	313.002106	413.273418
Поверхностная плотность, г/м2	693.225017	1399.542362
Модуль упругости при растяжении, ГПа	75.356612	82.682051
Прочность при растяжении, МПа	2767.193119	3848.436732
Потребление смолы, г/м2	257.481724	414.590628
Угол нашивки, град	90.000000	90.000000
Шаг нашивки	8.586293	14.440522
Плотность нашивки	64.944961	103.988901

Название показателя	Описание
Count	Количество элементов в столбце
Mean	Среднее арифметическое
Std	Среднее квадратичное отклонение
Min	Минимальное значение
Max	Максимальное значение
25%	Первый (25%) квартиль
50%	Медиана, второй (50%) квартиль
75%	Третий (75%) квартиль

Таким образом, вывели количество элементов (count), среднее арифметическое (mean), среднее квадратичное отклонение (std), минимальное (min) и максимальное (max) значение, первый (25%), второй (50%) и третий (75%) квартили по всему датасету. Дополнительно отмечаем, что данные в таблице определяются в разных единицах

измерения и имеется большой разброс в масштабе данных. Так же такую информацию можно получить с помощью методов df.mean(), df.median(), df.std(), df.min(), df.max() и т.д.

Проверям датасет на наличие дубликатов

```
df.duplicated()  
0      False  
1      False  
2      False  
3      False  
4      False  
...  
1018    False  
1019    False  
1020    False  
1021    False  
1022    False  
Length: 1023, dtype: bool  
df.duplicated().sum()  
np.int64(0)
```

Дублирующей информации в датасете не обнаружено

### Осуществляем исследование данных с помощью визуальных способов исследования (построения графиков)

Еще раз выводим информацию о первых и последних 5 строках датасета для наглядности

```
df.head()  
Соотношение матрица-наполнитель Плотность, кг/м3 модуль  
упругости, ГПа \  
0          1.857143        2030.0  
738.736842  
1          1.857143        2030.0  
738.736842  
2          1.857143        2030.0  
738.736842  
3          1.857143        2030.0  
738.736842  
4          2.771331        2030.0  
753.000000  
  
Количество отвердителя, м.% Содержание эпоксидных групп,%_2 \  
0          30.00          22.267857  
1          50.00          23.750000  
2          49.90          33.000000  
3          129.00         21.250000
```

4	111.86	22.267857		
0	Температура вспышки, С_2	Поверхностная плотность, г/м2 \ 100.000000 210.0		
1	284.615385	210.0		
2	284.615385	210.0		
3	300.000000	210.0		
4	284.615385	210.0		
0	Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа 70.0 3000.0		
1		3000.0		
2		3000.0		
3		3000.0		
4		3000.0		
0	Потребление смолы, г/м2	Угол нашивки, град	Шаг нашивки	Плотность нашивки
57.0	220.0		0	4.0
1	220.0		0	4.0
60.0	220.0		0	4.0
70.0	220.0		0	5.0
47.0	220.0		0	5.0
4	220.0		0	5.0
57.0				
df.tail()				
1018	Соотношение матрица-наполнитель	Плотность, кг/м3 \ 2.271346 1952.087902		
1019		3.444022 2050.089171		
1020		3.280604 1972.372865		
1021		3.705351 2066.799773		
1022		3.808020 1890.413468		
1018	модуль упругости, ГПа	Количество отвердителя, м.% \ 912.855545 86.992183		
1019		444.732634 145.981978		
1020		416.836524 110.533477		
1021		741.475517 141.397963		
1022		417.316232 129.183416		

	Содержание эпоксидных групп, %_2	Температура вспышки, С_2 \
1018	20.123249	324.774576
1019	19.599769	254.215401
1020	23.957502	248.423047
1021	19.246945	275.779840
1022	27.474763	300.952708
\	Поверхностная плотность, г/м2	Модуль упругости при растяжении, ГПа \
1018	209.198700	
73.090961		
1019	350.660830	
72.920827		
1020	740.142791	
74.734344		
1021	641.468152	
74.042708		
1022	758.747882	
74.309704		
	Прочность при растяжении, МПа	Потребление смолы, г/м2 \
1018	2387.292495	125.007669
1019	2360.392784	117.730099
1020	2662.906040	236.606764
1021	2071.715856	197.126067
1022	2856.328932	194.754342
Угол нашивки, град	Шаг нашивки	Плотность нашивки
1018	90	9.076380
1019	90	10.565614
1020	90	4.161154
1021	90	6.313201
1022	90	6.078902

Строим графики гистограмм распределения значений по каждому столбцу (переменной) в датасете.

Гистограммы используются для изучения распределения данных, выявления тенденций и закономерностей. Дополнительно на графики налагаем кривую оценки плотности ядра (KDE), которая помогает визуализировать распределение точек данных в наборе данных и выявлять закономерности, кластеры и аномалии в данных.

```
# Строим холст для нанесения 13 графиков (по количеству столбцов).
a = 5 # количество графиков в строке холста
b = 5 # количество график в столбце холста
c = 1 # порядковый номер графика

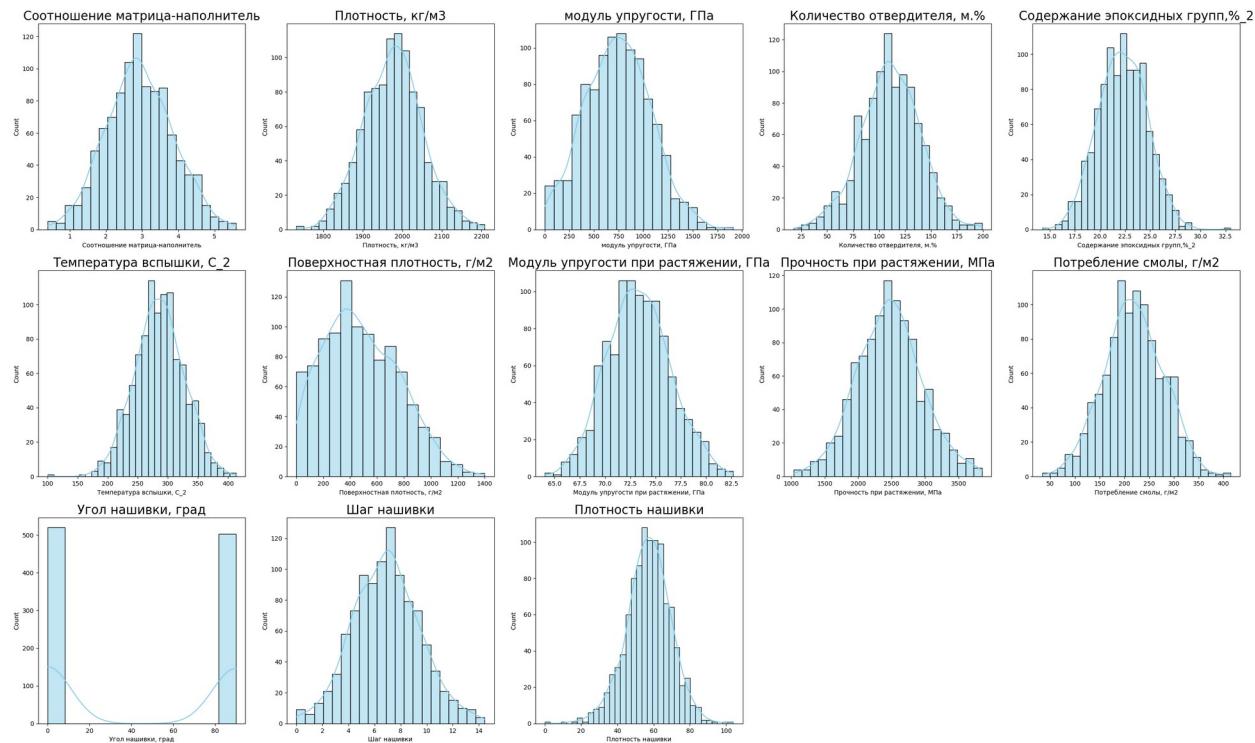
plt.figure(figsize = (35,35))
plt.suptitle('Гистограммы переменных', fontsize = 25)
```

```

for i in df:
    plt.subplot(a, b, c)
    sns.histplot(data=df[i], kde=True, color="skyblue")
    plt.title(i, size = 20)
    c = c + 1

```

Гистограммы переменных



Отмечаем, что распределение переменной "Угол нашивки, град" бимодально, т.е. содержит только два значения - 0 и 90. Отмечаем, что на гистограмме переменной "Поверхностная плотность, г/м<sup>2</sup>" наблюдается асимметрия вправо. По остальным переменным распределение напоминает нормальное.

Проверяем нормальность распределения данных

Проводим статистический тест Шапиро-Уилка, который позволяет проверить гипотезу о нормальности распределения данных. Используем модуль stats.

```

#Проверка на нормальность в Scipy
for i in list(df.columns.values):
    stat, p = stats.shapiro(df[i]) # тест Шапиро-Уилк
    print(i)
    print('Statistics=%.3f, p-value=%.3f' % (stat, p))

alpha = 0.05

```

```
if p > alpha:  
    print('Принять гипотезу о нормальности распределения\n')  
else:  
    print('Отклонить гипотезу о нормальности распределения\n')
```

Соотношение матрица-наполнитель  
Statistics=0.998, p-value=0.278  
Принять гипотезу о нормальности распределения

Плотность, кг/м3  
Statistics=0.999, p-value=0.550  
Принять гипотезу о нормальности распределения

модуль упругости, ГПа  
Statistics=0.996, p-value=0.007  
Отклонить гипотезу о нормальности распределения

Количество отвердителя, м.%  
Statistics=0.998, p-value=0.209  
Принять гипотезу о нормальности распределения

Содержание эпоксидных групп,%\_2  
Statistics=0.998, p-value=0.265  
Принять гипотезу о нормальности распределения

Температура вспышки, С\_2  
Statistics=0.998, p-value=0.283  
Принять гипотезу о нормальности распределения

Поверхностная плотность, г/м2  
Statistics=0.978, p-value=0.000  
Отклонить гипотезу о нормальности распределения

Модуль упругости при растяжении, ГПа  
Statistics=0.998, p-value=0.174  
Принять гипотезу о нормальности распределения

Прочность при растяжении, МПа  
Statistics=0.998, p-value=0.221  
Принять гипотезу о нормальности распределения

Потребление смолы, г/м2  
Statistics=0.999, p-value=0.708  
Принять гипотезу о нормальности распределения

Угол нашивки, град  
Statistics=0.636, p-value=0.000  
Отклонить гипотезу о нормальности распределения

Шаг нашивки

```
Statistics=0.998, p-value=0.176
Принять гипотезу о нормальности распределения
```

```
Плотность нашивки
Statistics=0.993, p-value=0.000
Отклонить гипотезу о нормальности распределения
```

На основании гистограмм распределения и статистического теста Шапиро-Уилка делаем выводы о том, что значения переменных "Соотношение матрица-наполнитель", "Плотность, кг/м<sup>3</sup>", "Количество отвердителя, м.%", "Содержание эпоксидных групп,%\_2", "Температура вспышки, С\_2", "Модуль упругости при растяжении, ГПа", "Прочность при растяжении, МПа", "Потребление смолы, г/м<sup>2</sup>", "Шаг нашивки" распределены нормально.

Значения переменных "модуль упругости, ГПа", "Поверхностная плотность, г/м<sup>2</sup>", "Угол нашивки, град", "Плотность нашивки" распределены не нормально.

Строим диаграммы размаха (боксплот) по каждому столбцу (переменной) в датасете

Диаграмма размаха — это статистический инструмент для визуализации распределения данных. Она выглядит как прямоугольник с линиями по краям, которые указывают на минимальные и максимальные значения. Диаграммы размаха позволяют сделать наблюдения о таких о ключевых значениях, например как:

- средний показатель, медиана 25го перцентиля и так далее;
- наличие выбросов и каковы их значения;
- о симметричности распределения данных, группировка данных;
- направления смещения данных.

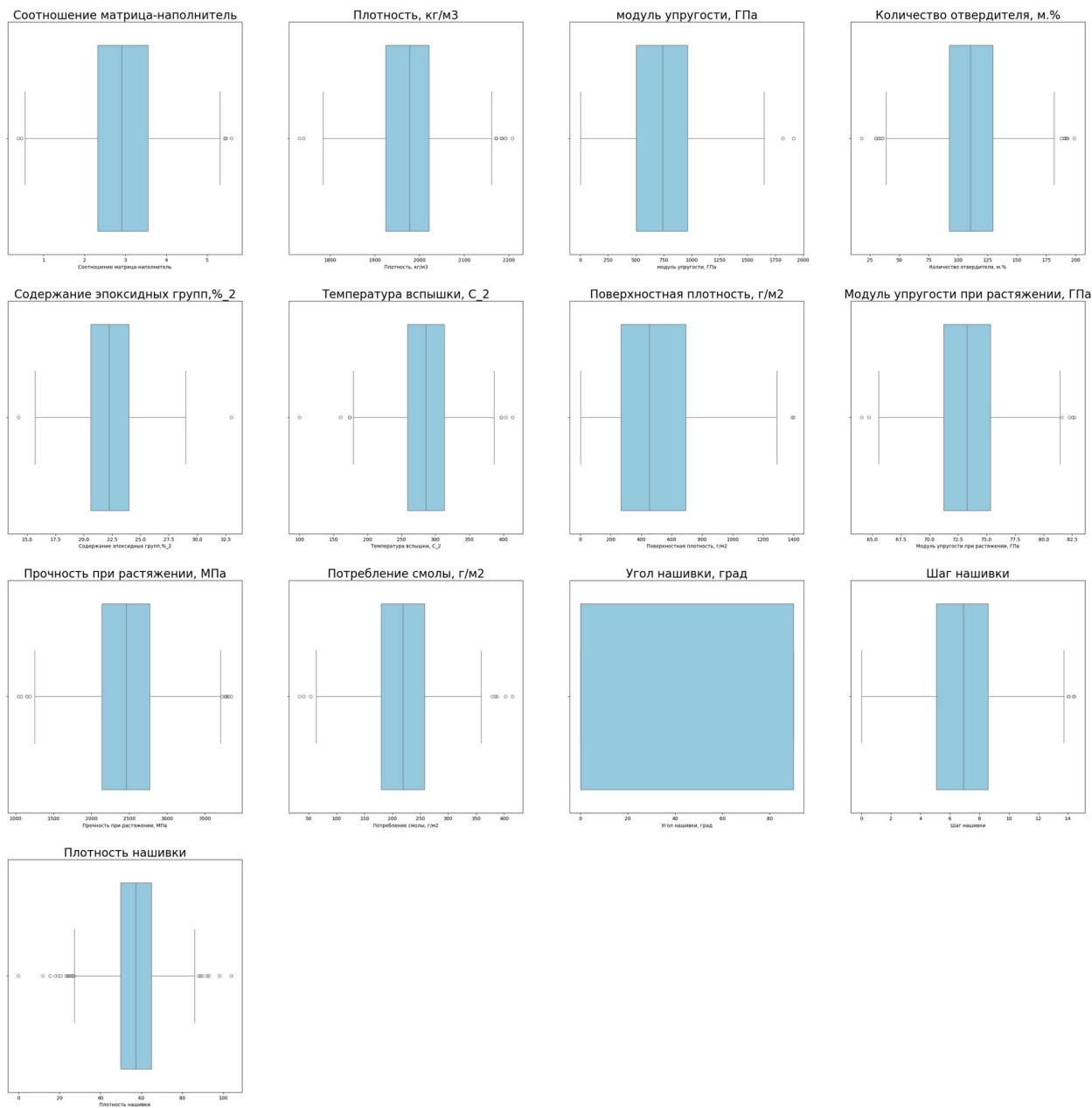
```
# Строим холст для нанесения 13 графиков (по количеству столбцов).
```

```
a = 4 # количество графиков в строке холста
b = 4 # количество график в столбце холста
c = 1 # порядковый номер графика

plt.figure(figsize = (40,40))
plt.suptitle('Диаграммы размаха', fontsize = 30)

for i in df:
    plt.subplot(a, b, c)
    sns.boxplot(data = df[i], color = "skyblue", orient='h')
    plt.title(i, size = 24)
    c = c + 1
```

### Диаграммы размаха



Обнаруживаем, что в каждой переменной имеются выбросы, за исключением переменной - "Угол нашивки". Переменная "Угол нашивки" является бимодальной и содержит только значения 0 и 90.

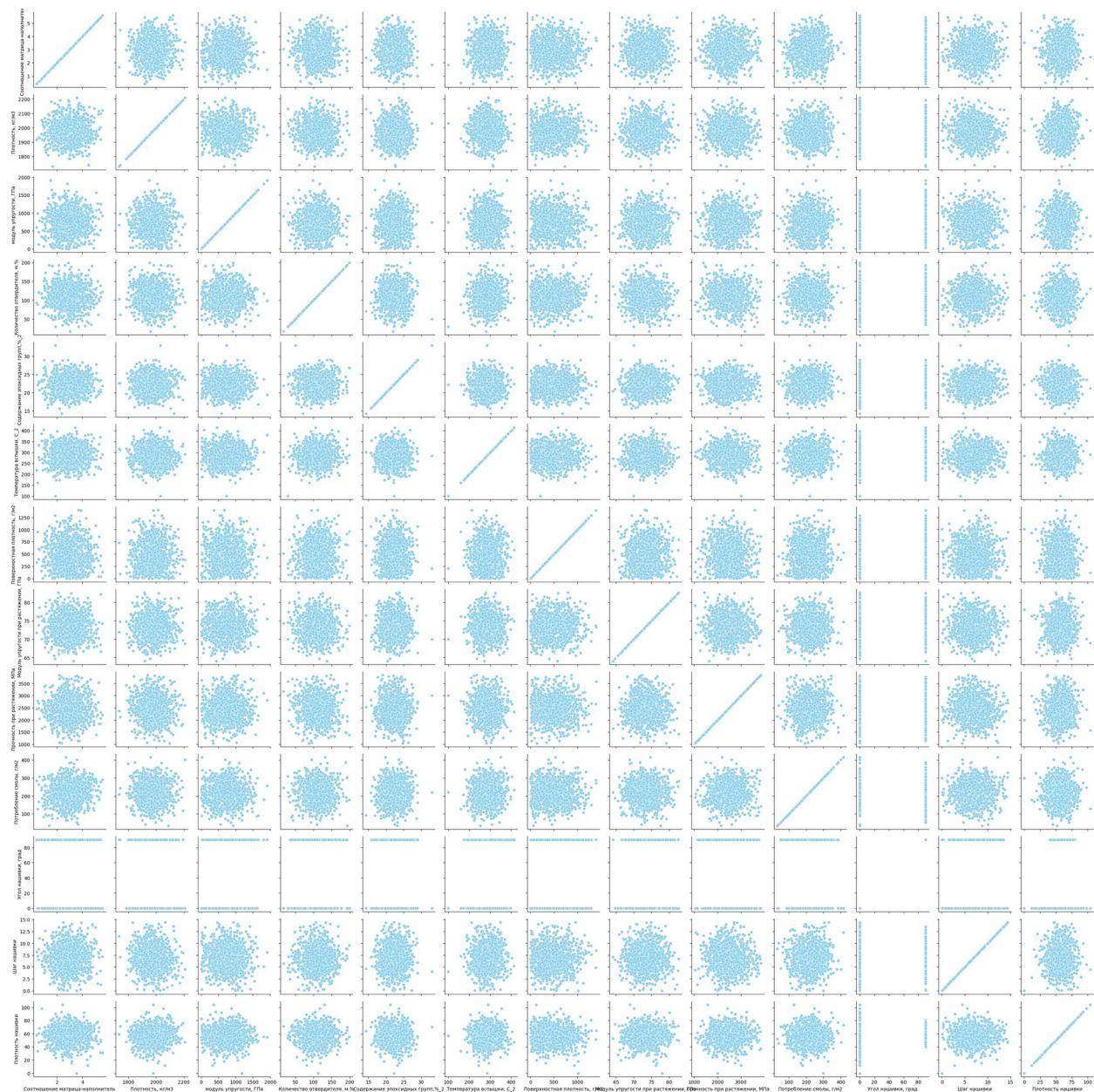
Строим парные диаграммы рассеяния точек

Отображаемые на диаграммах рассеяния паттерны позволяют увидеть разные типы корреляции. Среди них: положительная (оба значения увеличиваются), отрицательная (одно значение увеличивается, в то время как второе уменьшается), нулевая (отсутствие

корреляции), линейная, экспоненциальная и подковообразная. Сила корреляции определяется по тому, насколько близко расположены друг от друга точки на графике. Точки, которые значительно удалены от общего кластера точек, называются выбросами.

```
# для отрисовки диаграмм рассеяния используем модуль seaborn
parn_gr_rass = sns.PairGrid(df[df.columns])
parn_gr_rass.map(sns.scatterplot, color="skyblue")
plt.show

<function matplotlib.pyplot.show(close=None, block=None)>
```



Обнаруживаем отсутствие как положительной, так и отрицательной корреляции между переменными. Из диаграмм рассеяния можно сделать вывод об отсутствии корреляции

между переменными. Так же на диаграммах рассеяния видны размещенные точки (выбросы)

Дополнительно проверяем наличие корреляции между переменными. Используем метод pd.DataFrame.corr() и вычисляем значение коэффициента корреляции значений переменных в столбцах.

```
df.corr()
```

Соотношение матрица-наполнитель	
\	1.000000
Соотношение матрица-наполнитель	
Плотность, кг/м3	0.003841
модуль упругости, ГПа	0.031700
Количество отвердителя, м.%	-0.006445
Содержание эпоксидных групп,%_2	0.019766
Температура вспышки, С_2	-0.004776
Поверхностная плотность, г/м2	-0.006272
Модуль упругости при растяжении, ГПа	-0.008411
Прочность при растяжении, МПа	0.024148
Потребление смолы, г/м2	0.072531
Угол нашивки, град	-0.031073
Шаг нашивки	0.036437
Плотность нашивки	-0.004652

	Плотность, кг/м3	модуль
упругости, ГПа \		
Соотношение матрица-наполнитель	0.003841	
0.031700		
Плотность, кг/м3	1.000000	-
0.009647		
модуль упругости, ГПа	-0.009647	
1.000000		
Количество отвердителя, м.%	-0.035911	
0.024049		
Содержание эпоксидных групп,%_2	-0.008278	-
0.006804		
Температура вспышки, С_2	-0.020695	

0.031174	
Поверхностная плотность, г/м2	0.044930 -
0.005306	
Модуль упругости при растяжении, ГПа	-0.017602
0.023267	
Прочность при растяжении, МПа	-0.069981
0.041868	
Потребление смолы, г/м2	-0.015937
0.001840	
Угол нашивки, град	-0.068474 -
0.025417	
Шаг нашивки	-0.061015 -
0.009875	
Плотность нашивки	0.080304
0.056346	

	Количество отвердителя, м.% \
Соотношение матрица-наполнитель	-0.006445
Плотность, кг/м3	-0.035911
модуль упругости, ГПа	0.024049
Количество отвердителя, м.%	1.000000
Содержание эпоксидных групп,%_2	-0.000684
Температура вспышки, С_2	0.095193
Поверхностная плотность, г/м2	0.055198
Модуль упругости при растяжении, ГПа	-0.065929
Прочность при растяжении, МПа	-0.075375
Потребление смолы, г/м2	0.007446
Угол нашивки, град	0.038570
Шаг нашивки	0.014887
Плотность нашивки	0.017248

	Содержание эпоксидных групп,%_2 \
\ Соотношение матрица-наполнитель	0.019766
Плотность, кг/м3	-0.008278
модуль упругости, ГПа	-0.006804
Количество отвердителя, м.%	-0.000684
Содержание эпоксидных групп,%_2	1.000000
Температура вспышки, С_2	-0.009769
Поверхностная плотность, г/м2	-0.012940
Модуль упругости при растяжении, ГПа	0.056828
Прочность при растяжении, МПа	-0.023899

Потребление смолы, г/м2	0.015165
Угол нашивки, град	0.008052
Шаг нашивки	0.003022
Плотность нашивки	-0.039073

Соотношение матрица-наполнитель	Температура вспышки, С_2 \
Плотность, кг/м3	-0.004776
модуль упругости, ГПа	-0.020695
Количество отвердителя, м.%	0.031174
Содержание эпоксидных групп,%_2	0.095193
Температура вспышки, С_2	-0.009769
Поверхностная плотность, г/м2	1.000000
Модуль упругости при растяжении, ГПа	0.020121
Прочность при растяжении, МПа	0.028414
Потребление смолы, г/м2	-0.031763
Угол нашивки, град	0.059954
Шаг нашивки	0.020695
Плотность нашивки	0.025795
	0.011391

Соотношение матрица-наполнитель	Поверхностная плотность, г/м2 \
Плотность, кг/м3	-0.006272
модуль упругости, ГПа	0.044930
Количество отвердителя, м.%	-0.005306
Содержание эпоксидных групп,%_2	0.055198
Температура вспышки, С_2	-0.012940
Поверхностная плотность, г/м2	0.020121
Модуль упругости при растяжении, ГПа	1.000000
Прочность при растяжении, МПа	0.036702
Потребление смолы, г/м2	-0.003210
Угол нашивки, град	0.015692
Шаг нашивки	0.052299
Плотность нашивки	0.038332
	-0.049923

ГПа \	Модуль упругости при растяжении,
Соотношение матрица-наполнитель	-
0.008411	-
Плотность, кг/м3	-
0.017602	-
модуль упругости, ГПа	-
0.023267	-
Количество отвердителя, м.%	-
0.065929	-

Содержание эпоксидных групп,%\_2

0.056828

Температура вспышки, С\_2

0.028414

Поверхностная плотность, г/м2

0.036702

Модуль упругости при растяжении, ГПа

1.000000

Прочность при растяжении, МПа

0.009009

Потребление смолы, г/м2

0.050938

Угол нашивки, град

0.023003

Шаг нашивки

0.029468

Плотность нашивки

0.006476

Прочность при растяжении, МПа \

Соотношение матрица-наполнитель 0.024148

Плотность, кг/м3 -0.069981

модуль упругости, ГПа 0.041868

Количество отвердителя, м.% -0.075375

Содержание эпоксидных групп,%\_2 -0.023899

Температура вспышки, С\_2 -0.031763

Поверхностная плотность, г/м2 -0.003210

Модуль упругости при растяжении, ГПа -0.009009

Прочность при растяжении, МПа 1.000000

Потребление смолы, г/м2 0.028602

Угол нашивки, град 0.023398

Шаг нашивки -0.059547

Плотность нашивки 0.019604

Потребление смолы, г/м2 \

Соотношение матрица-наполнитель 0.072531

Плотность, кг/м3 -0.015937

модуль упругости, ГПа 0.001840

Количество отвердителя, м.% 0.007446

Содержание эпоксидных групп,%\_2 0.015165

Температура вспышки, С\_2 0.059954

Поверхностная плотность, г/м2 0.015692

Модуль упругости при растяжении, ГПа 0.050938

Прочность при растяжении, МПа 0.028602

Потребление смолы, г/м2 1.000000

Угол нашивки, град -0.015334

Шаг нашивки 0.013394

Плотность нашивки 0.012239

Угол нашивки, град Шаг нашивки

\		
Соотношение матрица-наполнитель	-0.031073	0.036437
Плотность, кг/м3	-0.068474	-0.061015
модуль упругости, ГПа	-0.025417	-0.009875
Количество отвердителя, м.%	0.038570	0.014887
Содержание эпоксидных групп,%_2	0.008052	0.003022
Температура вспышки, С_2	0.020695	0.025795
Поверхностная плотность, г/м2	0.052299	0.038332
Модуль упругости при растяжении, ГПа	0.023003	-0.029468
Прочность при растяжении, МПа	0.023398	-0.059547
Потребление смолы, г/м2	-0.015334	0.013394
Угол нашивки, град	1.000000	0.023616
Шаг нашивки	0.023616	1.000000
Плотность нашивки	0.107947	0.003487
	Плотность нашивки	
Соотношение матрица-наполнитель	-0.004652	
Плотность, кг/м3	0.080304	
модуль упругости, ГПа	0.056346	
Количество отвердителя, м.%	0.017248	
Содержание эпоксидных групп,%_2	-0.039073	
Температура вспышки, С_2	0.011391	
Поверхностная плотность, г/м2	-0.049923	
Модуль упругости при растяжении, ГПа	0.006476	
Прочность при растяжении, МПа	0.019604	
Потребление смолы, г/м2	0.012239	
Угол нашивки, град	0.107947	
Шаг нашивки	0.003487	
Плотность нашивки	1.000000	

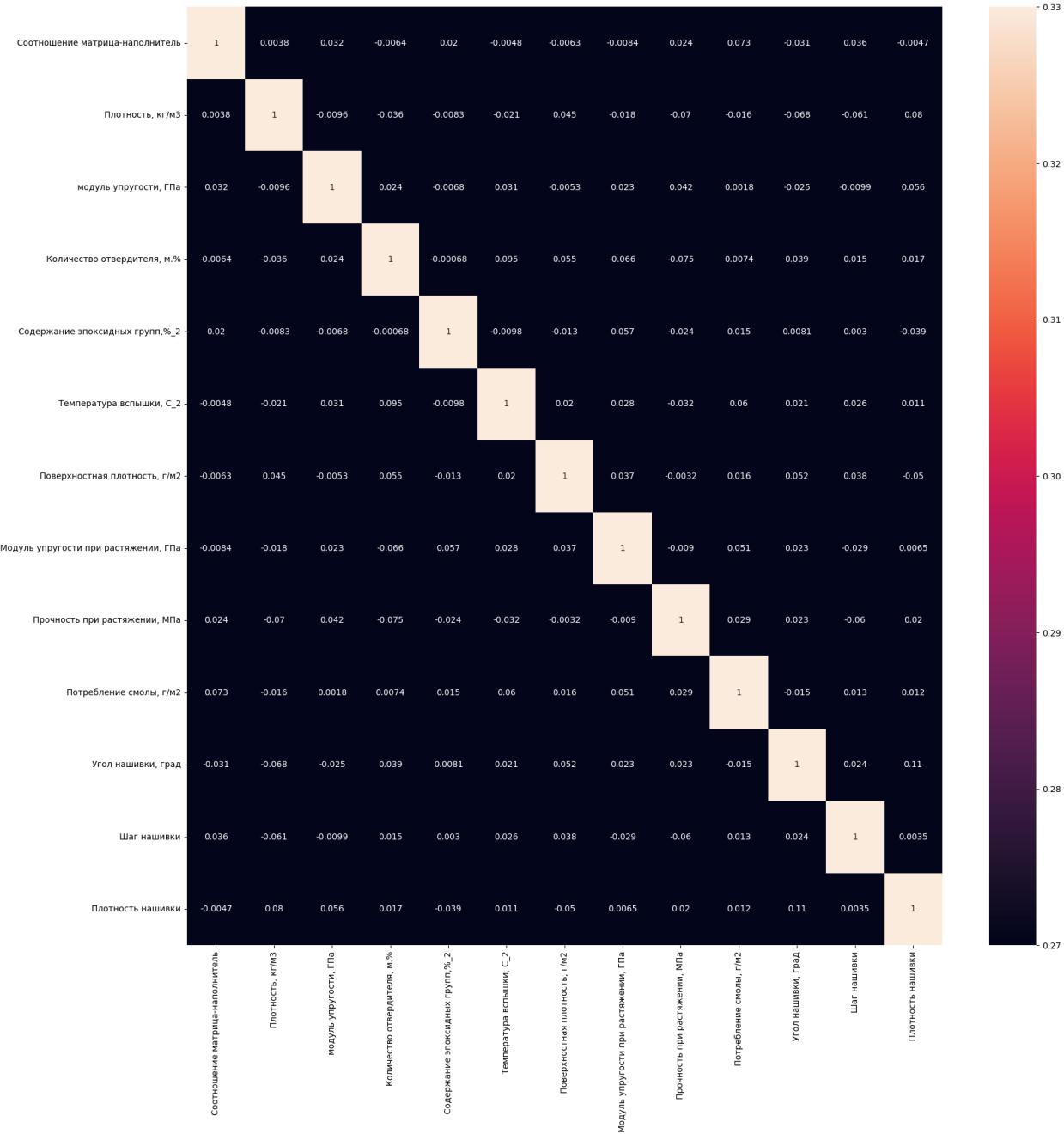
Корреляция - это статистическая взаимосвязь между двумя или более переменными. Она указывает на то, насколько изменения в одной переменной связаны с изменениями в другой. Если изменение одной переменной сопровождается систематическим изменением другой, то говорят о наличии корреляции. Корреляция не обязательно означает причинно-следственную связь, то есть изменение одной переменной не обязательно является причиной изменения другой, они могут быть связаны через третью переменную или просто случайно.

Коэффициент корреляции: Для измерения силы и направления связи между переменными используется коэффициент корреляции, который обычно принимает значения от -1 до +1. Значение, близкое к +1 или -1, указывает на сильную корреляцию, а значение, близкое к 0, указывает на слабую или отсутствующую корреляцию.

Дополнительно строим тепловую карту попарной корреляции столбцов датасета.

```
#Отобразим тепловую матрицу для наглядности
fig = plt.figure(figsize = (20,20))
sns.heatmap(df.corr(), annot = True, vmax=0.3, vmin=-0.3)

<Axes: >
```



Отмечаем, что большинство значений коэффициента корреляции близко к 0, что указывает на отсутствие корреляции. Максимальное значение коэффициента корреляции составляет 0.11 между переменными "Плотность нашивки" и "Угол нашивки, град", что говорит о наличии очень слабой силе связи между переменными.

## Обработка данных датасета

```
df.head()
```

Соотношение матрица-наполнитель упругости, ГПа	\	Плотность, кг/м3	модуль
0		1.857143	2030.0
738.736842			
1		1.857143	2030.0
738.736842			
2		1.857143	2030.0
738.736842			
3		1.857143	2030.0
738.736842			
4		2.771331	2030.0
753.000000			

Количество отвердителя, м.%	\	Содержание эпоксидных групп,%	2	\
0		30.00	22.267857	
1		50.00	23.750000	
2		49.90	33.000000	
3		129.00	21.250000	
4		111.86	22.267857	

Температура вспышки, С	2	\	Поверхностная плотность, г/м2	\
0			100.000000	210.0
1			284.615385	210.0
2			284.615385	210.0
3			300.000000	210.0
4			284.615385	210.0

Модуль упругости при растяжении, ГПа	\	Прочность при растяжении, МПа
0		70.0
1		3000.0
2		70.0
3		3000.0
4		70.0
		3000.0

Потребление смолы, г/м2	Угол нашивки, град	Шаг нашивки	Плотность
нашивки			
0	220.0	0	4.0
57.0			
1	220.0	0	4.0
60.0			
2	220.0	0	4.0
70.0			
3	220.0	0	5.0
47.0			

4  
57.0

220.0

0

5.0

Обрабатываем значения переменной "Угол нашивки, град".

Как установлено ранее, переменная имеет значения 0 или 90 градусов, распределение значения отличается от нормального. Кроме того, учитываем следующую информацию. Угол нашивки в композитных материалах, или угол ориентации волокна, определяет направление волокон в слое композита. Он влияет на прочность и жесткость композитного материала, а также на его поведение при нагрузках. Угол нашивки (направления волокна) - это угол между направлением волокон в слое композитного материала и некоторым эталонным направлением (обычно осью X или L). Угол нашивки обычно указывается в градусах. Например, 0° означает, что волокна параллельны эталонной оси, а 90° - перпендикулярны. Влияние на характеристики композита:

Прочность. Ориентация волокон влияет на прочность композита в разных направлениях. Композиты с ориентацией волокон, параллельной нагрузке, обычно имеют большую прочность на растяжение, чем те, где волокна перпендикулярны нагрузке.

Жесткость. Аналогично прочности, угол ориентации волокон влияет на жесткость композита. Ориентация волокон, параллельная нагрузке, обычно обеспечивает большую жесткость. Устойчивость к трещинам. Угол ориентации волокон может влиять на скорость распространения трещин в композите. В некоторых случаях, использование углов 45° или -45° может снижать риск быстрого разрушения. Вес. Композитные материалы могут быть легче и иметь при этом лучшие характеристики, чем традиционные металлы. Это достигается за счет оптимального выбора ориентации волокон. Примеры: 0°: Волокна параллельны оси нагрузки, обеспечивая максимальную прочность на растяжение в этом направлении. 90°: Волокна перпендикулярны оси нагрузки, обеспечивая высокую прочность на сжатие и сдвиг в этом направлении. 45°/-45°: Ориентация волокон под углом 45° и -45° относительно оси нагрузки может повысить сопротивление композита к сдвиговым нагрузкам и снизить риск хрупкого разрушения.

Необходимо отметить, что категориальные переменные — это те переменные, которые выражены ограниченным диапазоном значений. В нашем случае "Угол нашивки, град" выражен только 2 значениями - 0 или 90. В данном случае "Угол нашивки, град" необходимо отнести к номинальной переменной. Т.к. установить порядок и ранжирование значений не возможно.

При этом необходимо отметить, что угол нашивки будет влиять на свойство композитного материала в зависимости от ориентации волокон. При этом из датасета не ясно в каком направлении применялась сила (перпендикулярно, параллельно или под иным углом) при замерах значений в датасете.

Кроме того, для того, чтобы оценить свойства композитного материала в зависимости от оси направления силы по отношению к углу нашивки необходимо проводить несколько испытаний с целью понимания на сколько угол нашивки влияет на отдельные свойства композитного материала. Из датасета мы видим что в таблице № 2 было представлено по 520 значений угла нашивки, но исследование уникальных значений показало, что мы не можем определить что происходило 2 замера одного и того же композитного материала в зависимости от направления угла нашивки. Кроме того, мы не можем определить направление оси применения силы при замерах. Так же неясно направление применение

силы которое необходимо учитывать при обучении алгоритма машинного обучения, который будет определять значения:

- Модуль упругости при растяжении, ГПа;
- Прочность при растяжении, МПа.

Таким образом, итоговые значения композитного материала могут изменяться в зависимости от направления применения силы по отношению к углу нашивки. Но ввиду вышеуказанного мы не можем ранжировать значения угла нашивки или говорить о том, что значение 90 лучше чем значение 0, или наоборот.

В связи с чем мы кодируем переменную "Угол нашивки, град" как категориальную с использованием OneHotEncoder. OneHotEncoder - это метод, используемый в машинном обучении для представления категориальных данных в виде числовых данных. Он преобразует каждую категорию в двоичный вектор, где только один элемент "включен" (1), а остальные "выключены" (0). Этот метод имеет решающее значение для алгоритмов, требующих числового ввода, и помогает предотвратить неправильное толкование порядковых отношений между категориями.

```
# Осуществляем кодирование переменной "Угол нашивки, град"
ohe = OneHotEncoder(sparse_output=False)
ohe.fit_transform(df[['Угол нашивки, град']])
ohe.get_feature_names_out()
df[['Угол нашивки, град_0', 'Угол нашивки, град_90']] =
ohe.fit_transform(df[['Угол нашивки, град']])
df.drop('Угол нашивки, град', axis=1, inplace=True)

# Проверяем внесенные изменения в датафрейм
df.head()
```

	Соотношение матрица-наполнитель	Плотность, кг/м3	модуль упругости, ГПа
0	1.857143	2030.0	738.736842
1	1.857143	2030.0	738.736842
2	1.857143	2030.0	738.736842
3	1.857143	2030.0	738.736842
4	2.771331	2030.0	753.000000

	Количество отвердителя, м.%	Содержание эпоксидных групп,%_2
0	30.00	22.267857
1	50.00	23.750000
2	49.90	33.000000
3	129.00	21.250000
4	111.86	22.267857

	Температура вспышки, С_2	Поверхностная плотность, г/м2
--	--------------------------	-------------------------------

0	100.000000	210.0
1	284.615385	210.0
2	284.615385	210.0
3	300.000000	210.0
4	284.615385	210.0
Модуль упругости при растяжении, ГПа		Прочность при растяжении, МПа
\		
0	70.0	3000.0
1	70.0	3000.0
2	70.0	3000.0
3	70.0	3000.0
4	70.0	3000.0
Потребление смолы, г/м2		Шаг нашивки
0	220.0	4.0
1	220.0	4.0
2	220.0	4.0
3	220.0	5.0
4	220.0	5.0
Плотность нашивки		\
0		57.0
1		60.0
2		70.0
3		47.0
4		57.0
Угол нашивки, град_0		Угол нашивки, град_90
0	1.0	0.0
1	1.0	0.0
2	1.0	0.0
3	1.0	0.0
4	1.0	0.0

Температура вспышки

Датадфрейм содержит переменную "Температура вспышки, C\_2".

Температура вспышки – это минимальная температура, при которой вещество выделяет достаточно паров, чтобы при поднесении источника зажигания произошло воспламенение паров и их кратковременное горение, но без устойчивого горения после удаления источника зажигания. Это важный параметр для оценки пожарной опасности веществ, особенно при их транспортировке, хранении и использовании. Температура вспышки не оказывает влияния на прочность и упругость изготовленного композитного материала. В связи с чем принимаем решения исключить из датасета переменную "Температура вспышки, C\_2".

```
# исключаем переменную "Температура вспышки, C_2"
df.drop('Температура вспышки, C_2', axis=1, inplace=True)
```

```
# Проверяем датафрейм после удаления переменной  
df.head()
```

	Соотношение матрица-наполнитель упругости, ГПа	Плотность, кг/м3	модуль
0	738.736842	1.857143	2030.0
1	738.736842	1.857143	2030.0
2	738.736842	1.857143	2030.0
3	738.736842	1.857143	2030.0
4	753.000000	2.771331	2030.0

	Количество отвердителя, м.%	Содержание эпоксидных групп,%_2	\
0	30.00	22.267857	
1	50.00	23.750000	
2	49.90	33.000000	
3	129.00	21.250000	
4	111.86	22.267857	

	Поверхностная плотность, г/м2	Модуль упругости при растяжении, ГПа
0	210.0	70.0
1	210.0	70.0
2	210.0	70.0
3	210.0	70.0
4	210.0	70.0

	Прочность при растяжении, МПа	Потребление смолы, г/м2	Шаг нашивки
0	3000.0	220.0	4.0
1	3000.0	220.0	4.0
2	3000.0	220.0	4.0
3	3000.0	220.0	5.0
4	3000.0	220.0	5.0

	Плотность нашивки	Угол нашивки, град_0	Угол нашивки, град_90
0	57.0	1.0	0.0

1	60.0	1.0	0.0
2	70.0	1.0	0.0
3	47.0	1.0	0.0
4	57.0	1.0	0.0

## Работа с Эпоксидными группами и Смолами

В датафрейме присутствуют 2 переменные "Потребление смолы г/м<sup>2</sup>" и "Содержание эпоксидных групп, %\_2".

Зная, что эпоксидные группы – это структурные единицы в молекулах эпоксидных смол, которые обеспечивают их способность к отвердеванию под действием отвердителей. Они представляют собой трехчленный цикл, содержащий один атом кислорода, и также называются оксиранами или эпоксидами. Эти группы реагируют с отвердителями, образуя сшитые (сетчатые) полимеры, которые отличаются высокой прочностью и устойчивостью к различным воздействиям.

Таким образом, несмотря на отсутствие корреляции между указанными переменными, мы видим, что "Содержание эпоксидных групп, %\_2" дано в процентном соотношении и знаем о том, что эпоксидные группы содержатся в смолах. Значения смол приведено граммах в столбце "Потребление смолы г/м<sup>2</sup>".

Поэтому принимаем, решение вычислить значение эпоксидных смол в содержащихся в смолах. Проводим вычисления и изменяем датасет фиксируем итоговое значение эпоксидных групп и количество смол (за вычетом содержащихся в них эпоксидных групп).

```
df['Содержание эпоксидных групп, г'] = df['Потребление смолы, г/м2'] * 
(df['Содержание эпоксидных групп,%_2'] / 100)

df.drop('Содержание эпоксидных групп,%_2', axis=1, inplace=True)

df['Потребление смолы, г/м2 без ЭГ'] = df['Потребление смолы, г/м2'] - 
df['Содержание эпоксидных групп, г']

df.drop('Потребление смолы, г/м2', axis=1, inplace=True)

# проверяем верность внесенных изменений
df.head()
```

Соотношение матрица-наполнитель упругости, ГПа \	Плотность, кг/м3	модуль
0 738.736842	1.857143	2030.0
1 738.736842	1.857143	2030.0
2 738.736842	1.857143	2030.0
3 738.736842	1.857143	2030.0
4 753.000000	2.771331	2030.0

	Количество отвердителя, м.%	Поверхностная плотность, г/м2	\
0	30.00	210.0	
1	50.00	210.0	
2	49.90	210.0	
3	129.00	210.0	
4	111.86	210.0	
	Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа	
\			
0	70.0	3000.0	
1	70.0	3000.0	
2	70.0	3000.0	
3	70.0	3000.0	
4	70.0	3000.0	
	Шаг нашивки	Плотность нашивки	Угол нашивки, град_0
\			
0	4.0	57.0	1.0
1	4.0	60.0	1.0
2	4.0	70.0	1.0
3	5.0	47.0	1.0
4	5.0	57.0	1.0
	Угол нашивки, град_90	Содержание эпоксидных групп, г	\
\			
0	0.0	48.989286	
1	0.0	52.250000	
2	0.0	72.600000	
3	0.0	46.750000	
4	0.0	48.989286	
	Потребление смолы, г/м2 без ЭГ		
\			
0	171.010714		
1	167.750000		
2	147.400000		
3	173.250000		
4	171.010714		

Таким образом, добавлены 2 новых переменных "Содержание эпоксидных групп, г" и "Потребление смолы, г/м2 без ЭГ". Так же понимаем, что существует связь отвердителя с взаимосвязью смолы и эпоксидных групп. Так как отвердитель и эпоксидные группы напрямую влияют на некоторые свойства композитных материалов. Так же существует сложности с трактованием значений отвердителя ввиду того, что он выражен в м.%. Кроме того, можем предположить, что данные переменные также могут прямо влиять на "плотность кг,м3". Но для формирования каких-либо выводом целесообразно получить

консультацию у эксперта, так же желательно знать вид эпоксидных групп, смол и отвердителя, т.к. это напрямую влияет на характеристики.

## Выбросы

Выброс — это экстремальные значения во входных данных, которые находятся далеко за пределами других наблюдений. Многие алгоритмы машинного обучения чувствительны к разбросу и распределению значений признаков обрабатываемых объектов.

Соответственно, выбросы во входных данных могут исказить и ввести в заблуждение процесс обучения алгоритмов машинного обучения, что приводит к увеличению времени обучения, снижению точности моделей и, в конечном итоге, к снижению результатов.

Даже до подготовки предсказательных моделей на основе обучающих данных выбросы могут приводить к ошибочным представлениям и в дальнейшем к ошибочной интерпретации собранных данных.

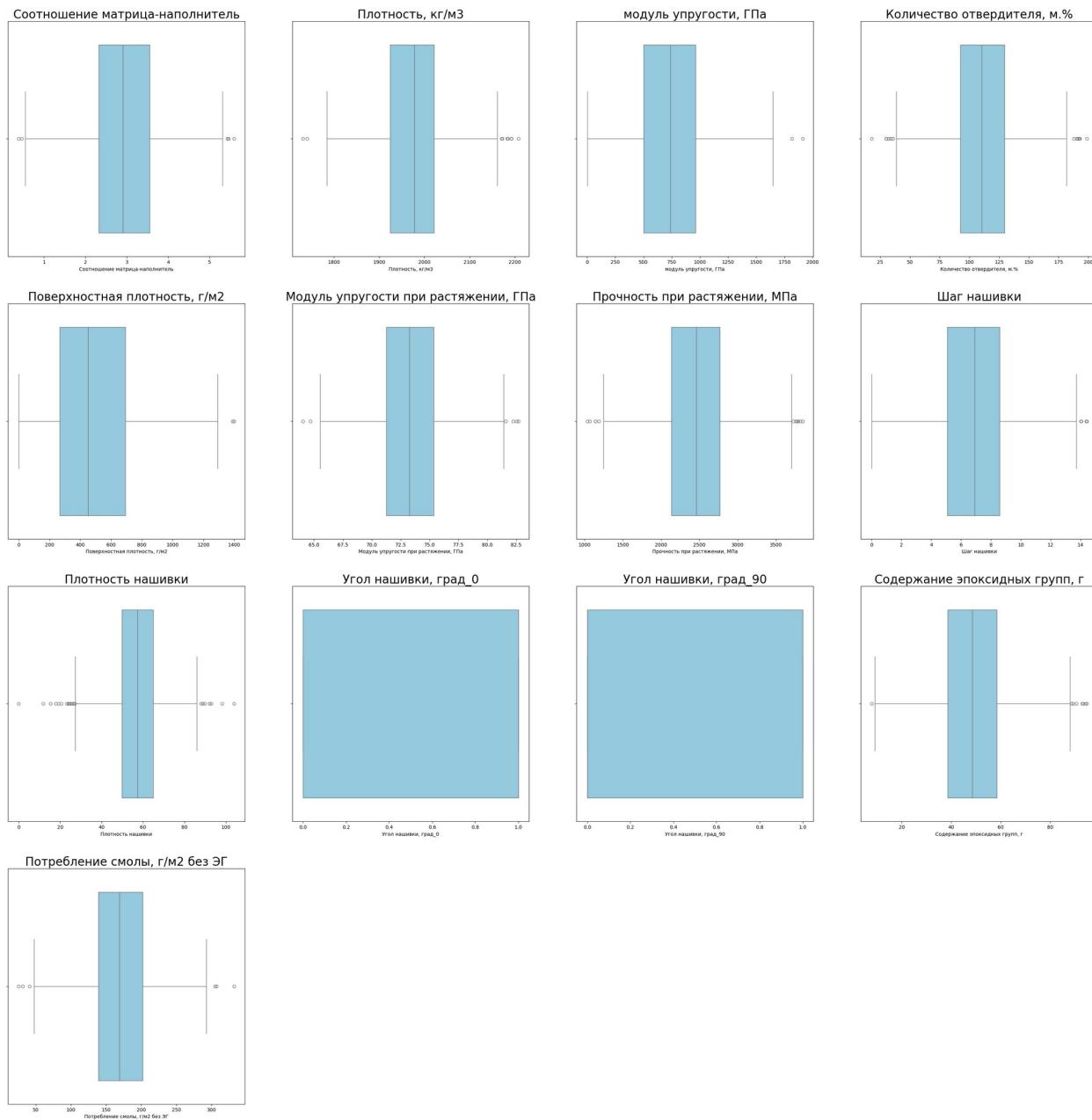
Ввиду того, что мы существенно изменили датафрейм. Отобразим повторно диаграммы размаха

```
a = 4 # количество графиков в строке холста
b = 4 # количество график в столбце холста
c = 1 # порядковый номер графика

plt.figure(figsize = (40,40))
plt.suptitle('Диаграммы размаха', fontsize = 30)

for i in df:
    plt.subplot(a, b, c)
    sns.boxplot(data=df[i], color="skyblue", orient="h")
    plt.title(i, size = 24)
    c = c + 1
```

### Диаграммы размаха



Диаграммы размаха указывают на наличие выбросов в переменных, за исключением кодированных переменных "Угол нашивки, град\_0" и "Угол нашивки\_град\_90".

Присваеваем выбросам значение NaN (пустое значение)

```
for col in df.columns:
    q75, q25 = np.percentile(df.loc[:, col], [75, 25])
    intr_qr = q75 - q25
```

```

max = q75 + (1.5 * intr_qr)
min = q25 - (1.5 * intr_qr)

df.loc[df[col] < min, col] = np.nan
df.loc[df[col] > max, col] = np.nan

```

Проверка количества выбросов (пустых значений)

```

df.isnull().sum().sum()
np.int64(89)

```

Количество выбросов составляет 89, что составляет 8.7% от общего количества значений в выборке. Принимаем решение об удалении выбросов.

```

df = df.dropna(axis=0)

df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 940 entries, 1 to 1022
Data columns (total 13 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Соотношение матрица-наполнитель    940 non-null   float64
 1   Плотность, кг/м3                  940 non-null   float64
 2   модуль упругости, ГПа            940 non-null   float64
 3   Количество отвердителя, м.%       940 non-null   float64
 4   Поверхностная плотность, г/м2     940 non-null   float64
 5   Модуль упругости при растяжении, ГПа 940 non-null   float64
 6   Прочность при растяжении, МПа      940 non-null   float64
 7   Шаг нашивки                     940 non-null   float64
 8   Плотность нашивки                940 non-null   float64
 9   Угол нашивки, град_0             940 non-null   float64
 10  Угол нашивки, град_90            940 non-null   float64
 11  Содержание эпоксидных групп, г     940 non-null   float64
 12  Потребление смолы, г/м2 без ЭГ     940 non-null   float64
dtypes: float64(13)
memory usage: 102.8 KB

```

Обнаруженные выбросы удалены. Повторно строим диаграммы размаха (боксплоты) с целью обнаружения выбросов

```

a = 4 # количество графиков в строке холста
b = 4 # количество график в столбце холста
c = 1 # порядковый номер графика

plt.figure(figsize = (40,40))

```

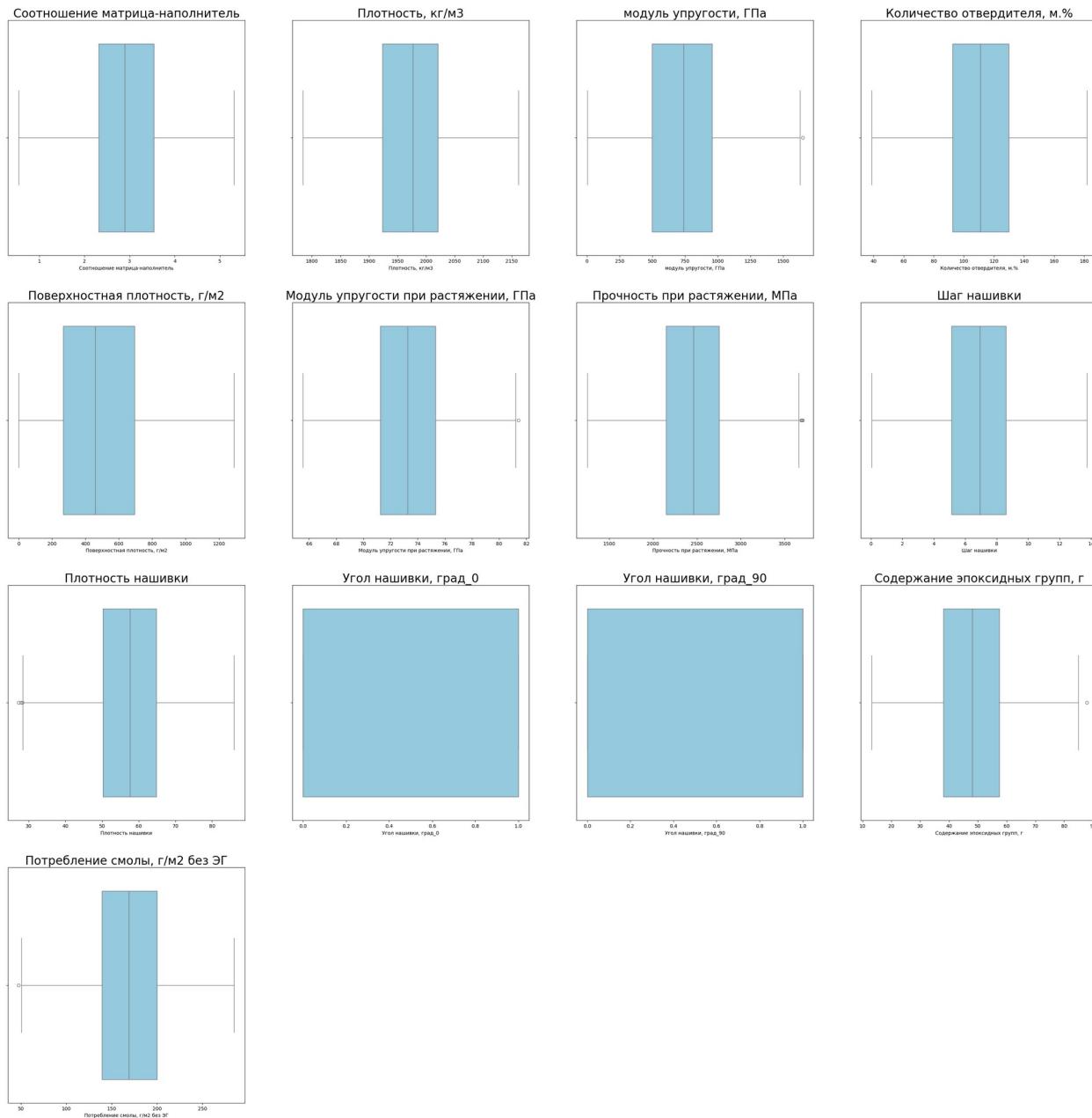
```

plt.suptitle('Диаграммы размаха', fontsize = 30)

for i in df:
    plt.subplot(a, b, c)
    sns.boxplot(data=df[i], color="skyblue", orient="h")
    plt.title(i, size = 24)
    c = c + 1

```

Диаграммы размаха



Некоторое количество выбросов осталось. Повторяем процедуру удаления выбросов.

```
for col in df.columns:  
    q75, q25 = np.percentile(df.loc[:, col], [75, 25])  
    intr_qr = q75 - q25  
  
    max = q75 + (1.5 * intr_qr)  
    min = q25 - (1.5 * intr_qr)  
  
    df.loc[df[col] < min, col] = np.nan  
    df.loc[df[col] > max, col] = np.nan  
  
df.isnull().sum().sum()  
  
np.int64(11)
```

Обнаружено еще 11 выбросов. Исключаем их.

```
df = df.dropna(axis=0)  
  
df.info()  
  
<class 'pandas.core.frame.DataFrame'>  
Index: 929 entries, 1 to 1022  
Data columns (total 13 columns):  
 #   Column           Non-Null Count  Dtype     
---  --  
 0   Соотношение матрица-наполнитель 929 non-null   float64  
 1   Плотность, кг/м3                 929 non-null   float64  
 2   модуль упругости, ГПа            929 non-null   float64  
 3   Количество отвердителя, м.%       929 non-null   float64  
 4   Поверхностная плотность, г/м2     929 non-null   float64  
 5   Модуль упругости при растяжении, ГПа 929 non-null   float64  
 6   Прочность при растяжении, МПа      929 non-null   float64  
 7   Шаг нашивки                     929 non-null   float64  
 8   Плотность нашивки                929 non-null   float64  
 9   Угол нашивки, град_0             929 non-null   float64  
 10  Угол нашивки, град_90            929 non-null   float64  
 11  Содержание эпоксидных групп, г    929 non-null   float64  
 12  Потребление смолы, г/м2 без ЭГ    929 non-null   float64  
dtypes: float64(13)  
memory usage: 101.6 KB
```

Проверяем выборку на наличие выбросов с помощью построения диаграммы размаха.

```
a = 4 # количество графиков в строке холста  
b = 4 # количество график в столбце холста  
c = 1 # порядковый номер графика  
  
plt.figure(figsize = (40,40))
```

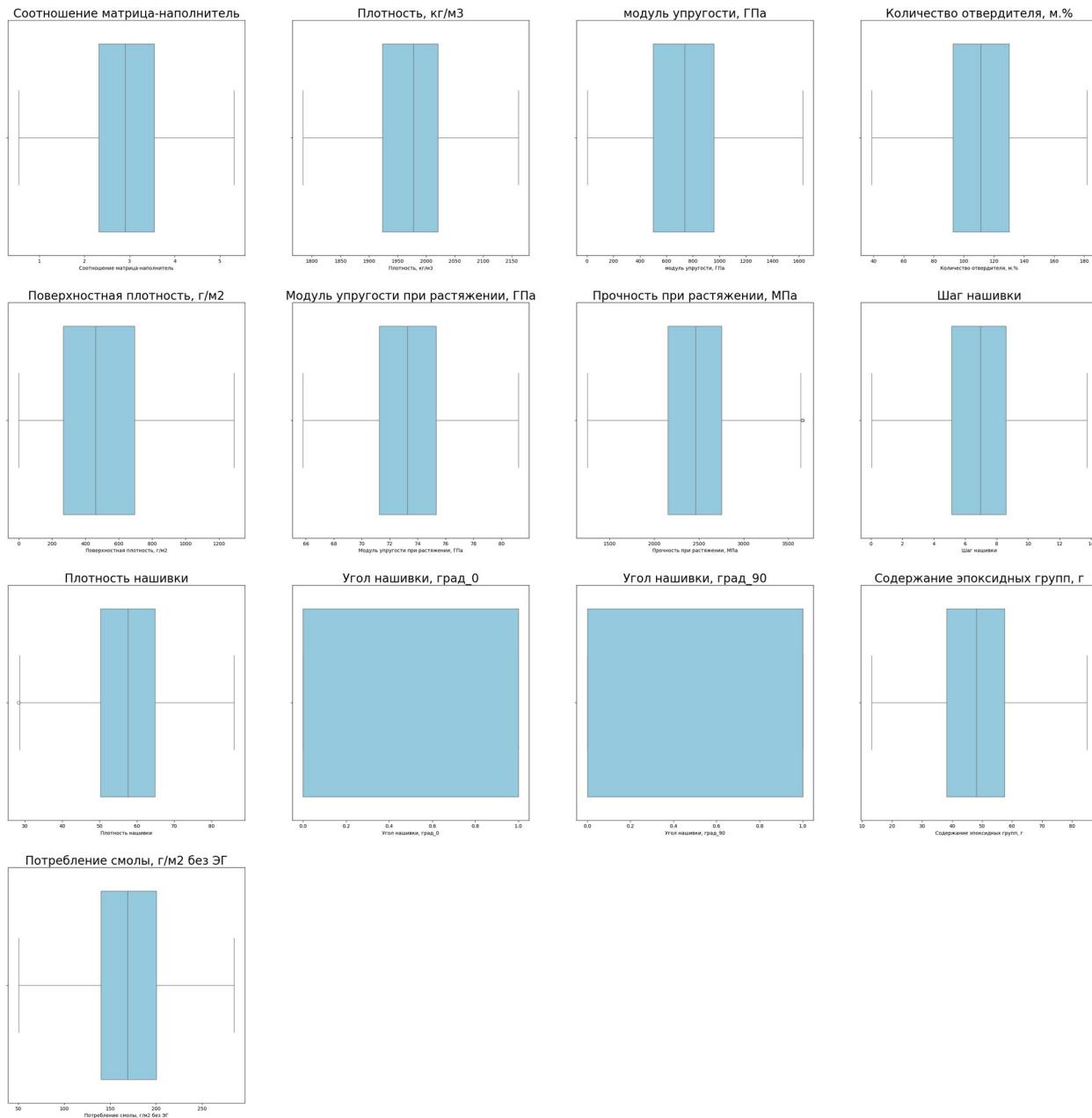
```

plt.suptitle('Диаграммы размаха', fontsize = 30)

for i in df:
    plt.subplot(a, b, c)
    sns.boxplot(data=df[i], color="skyblue", orient="h")
    plt.title(i, size = 24)
    c = c + 1

```

Диаграммы размаха



Некоторое количество выбросов осталось.

```
for col in df.columns:
    q75, q25 = np.percentile(df.loc[:, col], [75, 25])
    intr_qr = q75 - q25

    max = q75 + (1.5 * intr_qr)
    min = q25 - (1.5 * intr_qr)

    df.loc[df[col] < min, col] = np.nan
    df.loc[df[col] > max, col] = np.nan

df = df.dropna(axis=0)

df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 925 entries, 1 to 1022
Data columns (total 13 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Соотношение матрица-наполнитель  925 non-null   float64
 1   Плотность, кг/м3                 925 non-null   float64
 2   модуль упругости, ГПа            925 non-null   float64
 3   Количество отвердителя, м.%       925 non-null   float64
 4   Поверхностная плотность, г/м2     925 non-null   float64
 5   Модуль упругости при растяжении, ГПа 925 non-null   float64
 6   Прочность при растяжении, МПа      925 non-null   float64
 7   Шаг нашивки                     925 non-null   float64
 8   Плотность нашивки               925 non-null   float64
 9   Угол нашивки, град_0             925 non-null   float64
 10  Угол нашивки, град_90            925 non-null   float64
 11  Содержание эпоксидных групп, г     925 non-null   float64
 12  Потребление смолы, г/м2 без ЭГ      925 non-null   float64
dtypes: float64(13)
memory usage: 101.2 KB
```

Еще одно построение диаграммы размаха

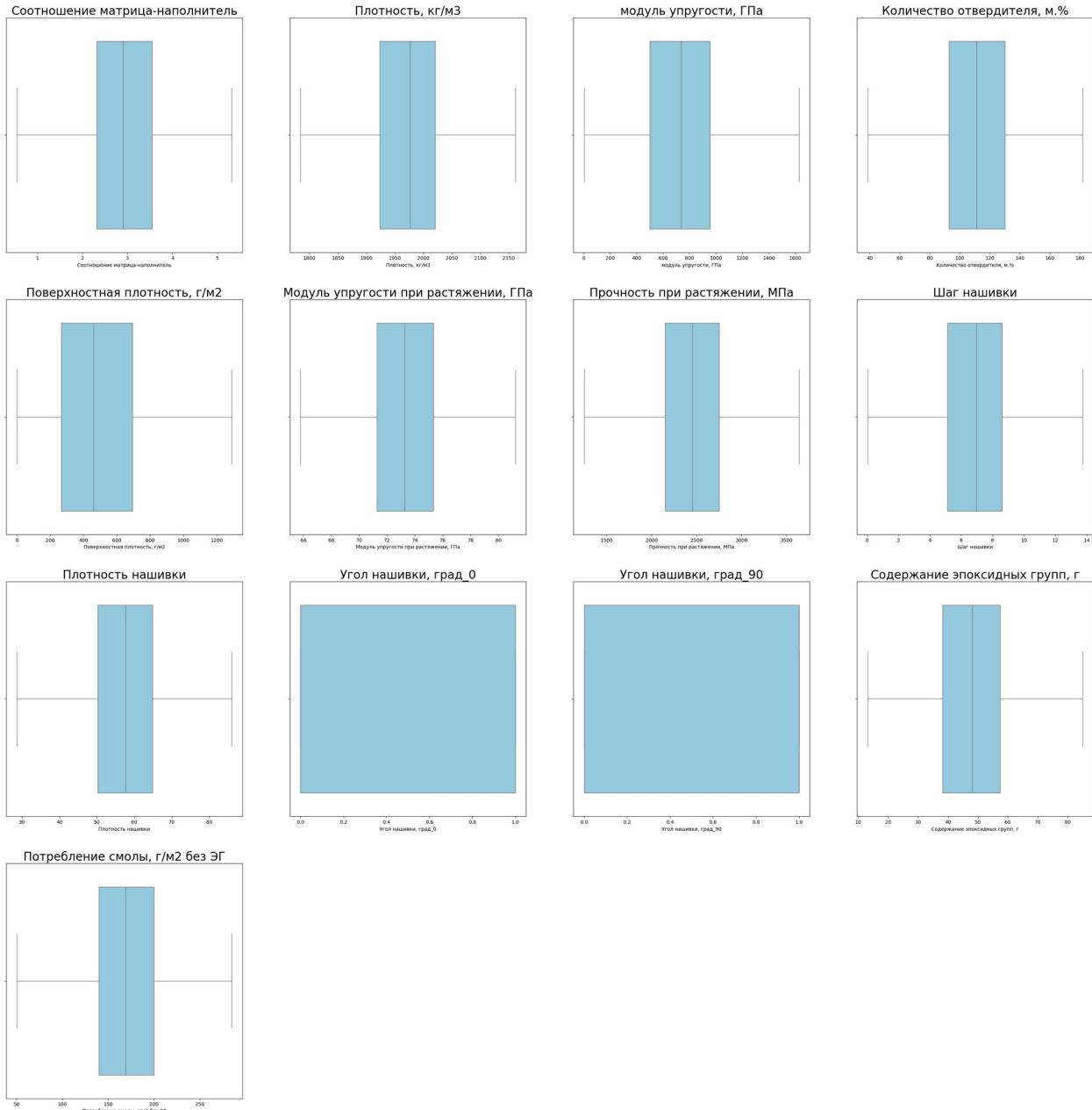
```
a = 4 # количество графиков в строке холста
b = 4 # количество график в столбце холста
c = 1 # порядковы номер графика

plt.figure(figsize = (40,40))
plt.suptitle('Диаграммы размаха', fontsize = 30)

for i in df:
    plt.subplot(a, b, c)
    sns.boxplot(data = df[i], color = "skyblue", orient='h')
```

```
plt.title(i, size = 24)
c = c + 1
```

Диаграммы размаха



Выбросы полностью исключены.

Масштабирование данных

# выводим первые 5 строк датафрейма
df.head()
Соотношение матрица-наполнитель \\\nупругости, ГПа Плотность, кг/м3 модуль
1 1.857143 2030.0
738.736842
2 1.857143 2030.0
738.736842
3 1.857143 2030.0
738.736842
4 2.771331 2030.0
753.000000
5 2.767918 2000.0
748.000000
Количество отвердителя, м.% Поверхностная плотность, г/м2 \\\n
1 50.00 210.0
2 49.90 210.0
3 129.00 210.0
4 111.86 210.0
5 111.86 210.0
Модуль упругости при растяжении, ГПа Прочность при растяжении, МПа \\\n
1 70.0 3000.0
2 70.0 3000.0
3 70.0 3000.0
4 70.0 3000.0
5 70.0 3000.0
Шаг нашивки Плотность нашивки Угол нашивки, град_0 \\\n
1 4.0 60.0 1.0
2 4.0 70.0 1.0
3 5.0 47.0 1.0
4 5.0 57.0 1.0
5 5.0 60.0 1.0
Угол нашивки, град_90 Содержание эпоксидных групп, г \\\n
1 0.0 52.250000
2 0.0 72.600000
3 0.0 46.750000
4 0.0 48.989286
5 0.0 48.989286

	Потребление смолы, г/м <sup>2</sup> без ЭГ	
1		167.750000
2		147.400000
3		173.250000
4		171.010714
5		171.010714

# выводим размеры датафрейма  
df.shape  
(925, 13)  
df.describe().T

		count	mean
std \			
Соотношение матрица-наполнитель		925.0	2.919068
Плотность, кг/м <sup>3</sup>		925.0	1974.319439
модуль упругости, ГПа		925.0	735.328695
Количество отвердителя, м.%		925.0	111.086088
Поверхностная плотность, г/м <sup>2</sup>		925.0	483.435186
Модуль упругости при растяжении, ГПа	925.0	73.304924	3.011295
Прочность при растяжении, МПа	925.0	2460.116851	450.916069
Шаг нашивки	925.0	6.918383	2.514591
Плотность нашивки	925.0	57.568830	11.065402
Угол нашивки, град_0	925.0	0.486486	0.500088
Угол нашивки, град_90	925.0	0.513514	0.500088
Содержание эпоксидных групп, г	925.0	48.129428	13.667150
Потребление смолы, г/м <sup>2</sup> без ЭГ	925.0	168.856094	44.576887

		min	25%
50% \			
Соотношение матрица-наполнитель		0.547391	2.314541
2.903233			
Плотность, кг/м <sup>3</sup>		1784.482245	1923.255135
1977.302956			
модуль упругости, ГПа		2.436909	498.519344
738.736842			

Количество отвердителя, м.%	38.668500	92.834720
111.157879		
Поверхностная плотность, г/м2	0.603740	267.140817
460.721126		
Модуль упругости при растяжении, ГПа	65.793845	71.279418
73.253725		
Прочность при растяжении, МПа	1250.392802	2149.974687
2456.395009		
Шаг нашивки	0.037639	5.117477
6.947322		
Плотность нашивки	28.661632	50.280645
57.615327		
Угол нашивки, град_0	0.000000	0.000000
0.000000		
Угол нашивки, град_90	0.000000	0.000000
1.000000		
Содержание эпоксидных групп, г	13.158372	38.137716
48.089010		
Потребление смолы, г/м2 без ЭГ	50.427004	139.606267
168.753112		
Соотношение матрица-наполнитель	75%	max
	3.546415	5.314144
Плотность, кг/м3	2020.828331	2161.565216
модуль упругости, ГПа	956.246864	1628.000000
Количество отвердителя, м.%	130.163998	181.828448
Поверхностная плотность, г/м2	694.589685	1291.340115
Модуль упругости при растяжении, ГПа	75.309657	81.203147
Прочность при растяжении, МПа	2751.235671	3636.892992
Шаг нашивки	8.606411	13.732404
Плотность нашивки	64.854936	86.012427
Угол нашивки, град_0	1.000000	1.000000
Угол нашивки, град_90	1.000000	1.000000
Содержание эпоксидных групп, г	57.358358	84.967714
Потребление смолы, г/м2 без ЭГ	199.907147	284.828567

В значениях переменных есть большой разброс в их масштабе. Отдельные значения могут быть равны 0 и/или близки к нему, а другие измеряются в тысячах. Ввиду того, что перед нами поставлена задача построения моделей машинного обучения и построение нейросети с целью прогноза значения зависимой переменной по значению независимых переменных отмечаем факт необходимости масштабирования данных.

Масштабирование данных — это процесс преобразования числовых признаков в наборе данных к одному масштабу, чтобы они имели одинаковый диапазон значений. Это необходимо, потому что многие алгоритмы машинного обучения чувствительны к разным масштабам признаков, и масштабирование помогает избежать ситуации, когда один признак доминирует над другими. Масштабирование применяется для: Улучшение работы алгоритмов. Многие алгоритмы, особенно те, что основаны на расстоянии (например, KNN, K-means, SVM) или градиентном спуске, работают лучше, когда признаки имеют одинаковый масштаб. Предотвращение доминирования признаков. Если один признак

имеет большой диапазон значений, а другой – маленький, то первый может оказывать большее влияние на модель, даже если его важность не выше. Масштабирование выравнивает вклад каждого признака. Оптимизация обучения. Масштабирование может ускорить процесс обучения модели и повысить ее точность.

Масштабирования данных будет произведено непосредственно перед созданием обучающей и тестовой выборок при построении моделей машинного обучения и модели нейросети (в частях 2 и 3 настоящего блокнота).

Анализ итогового датасета

```
df.head()
```

	Соотношение матрица-наполнитель упругости, ГПа	Плотность, кг/м3	модуль
--	---	------------------	--------

1	738.736842	1.857143	2030.0
2	738.736842	1.857143	2030.0
3	738.736842	1.857143	2030.0
4	753.000000	2.771331	2030.0
5	748.000000	2.767918	2000.0

	Количество отвердителя, м.%	Поверхностная плотность, г/м2	\
1	50.00	210.0	
2	49.90	210.0	
3	129.00	210.0	
4	111.86	210.0	
5	111.86	210.0	

	Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа	\
1	70.0	3000.0	
2	70.0	3000.0	
3	70.0	3000.0	
4	70.0	3000.0	
5	70.0	3000.0	

	Шаг нашивки	Плотность нашивки	Угол нашивки, град_0	\
1	4.0	60.0	1.0	
2	4.0	70.0	1.0	
3	5.0	47.0	1.0	
4	5.0	57.0	1.0	

5	5.0	60.0	1.0
1	Угол нашивки, град_90	Содержание эпоксидных групп, г \	
2	0.0	52.250000	
3	0.0	72.600000	
4	0.0	46.750000	
5	0.0	48.989286	
5	Потребление смолы, г/м2 без ЭГ	48.989286	
1		167.750000	
2		147.400000	
3		173.250000	
4		171.010714	
5		171.010714	

Наименование столбцов и описание их значений

Наименование столбца	Описание поля
Соотношение матрица-наполнитель	Матрица - это связующее вещество, в которое внедрен наполнитель (волокна, частицы, слои). Соотношение может быть выражено в процентах или в виде объемного или весового соотношения.
Плотность, кг/м3	Физическая величина, характеризующая массу вещества, содержащегося в единице объема. Выражен в килограмме на кубический метр.
модуль упругости, ГПа	Физическая величина, характеризующая способность твердого тела сопротивляться упругой деформации приложении к нему силы. Выражен в Гигапаскаль.
Количество отвердителя, м.%	Отвердитель – это химическое вещество, которое добавляется к лакокрасочным материалам, эпоксидным смолам и другим реакционноспособным олигомерам для ускорения или инициирования процесса отверждения (застывания, полимеризации).
Поверхностная плотность, г/м2	Величина, характеризующая массу вещества, приходящуюся на единицу площади. Выражен в грамм на метр кубический
Модуль упругости при растяжении, ГПа	Оценивает упругость жестких или твердых материалов, которая является соотношением между деформацией материала и силой, необходимой для его деформации. Выражен в Гигапаскаль.
Прочность при растяжении, МПа	Максимальное напряжение, которое материал может выдержать при растягивающей нагрузке

Наименование столбца	Описание поля
	до начала разрушения. Это важный показатель для оценки прочности и надежности материалов. Выражен в Мегапаскаль.
Шаг нашивки	Шаг нашивки
Плотность нашивки	Плотность нашивки
Угол нашивки, град_0	Угол нашивки равный 0 градусам
Угол нашивки, град_90	Угол нашивки равный 90 градусам
Содержание эпоксидных групп, г	Количество эпоксидных групп в смоле (в граммах). Эпоксидные группы – это структурные единицы в молекулах эпоксидных смол, которые обеспечивают их способность к отвердеванию под действием отвердителей.
Потребление смолы, г/м2 без ЭГ	Потребление смолы за вычетом эпоксидных групп в смоле (в граммах/м2)

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 925 entries, 1 to 1022
Data columns (total 13 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Соотношение матрица-наполнитель    925 non-null   float64
 1   Плотность, кг/м3                  925 non-null   float64
 2   модуль упругости, ГПа            925 non-null   float64
 3   Количество отвердителя, м.%       925 non-null   float64
 4   Поверхностная плотность, г/м2      925 non-null   float64
 5   Модуль упругости при растяжении, ГПа 925 non-null   float64
 6   Прочность при растяжении, МПа        925 non-null   float64
 7   Шаг нашивки                      925 non-null   float64
 8   Плотность нашивки                925 non-null   float64
 9   Угол нашивки, град_0             925 non-null   float64
 10  Угол нашивки, град_90            925 non-null   float64
 11  Содержание эпоксидных групп, г     925 non-null   float64
 12  Потребление смолы, г/м2 без ЭГ      925 non-null   float64
dtypes: float64(13)
memory usage: 101.2 KB
```

```
df.describe().T
```

	count	mean
std \		
Соотношение матрица-наполнитель	925.0	2.919068
Плотность, кг/м3	925.0	1974.319439
модуль упругости, ГПа	925.0	735.328695
	327.705942	

Количество отвердителя, м.%	925.0	111.086088	26.830806
Поверхностная плотность, г/м2	925.0	483.435186	279.103156
Модуль упругости при растяжении, ГПа	925.0	73.304924	3.011295
Прочность при растяжении, МПа	925.0	2460.116851	450.916069
Шаг нашивки	925.0	6.918383	2.514591
Плотность нашивки	925.0	57.568830	11.065402
Угол нашивки, град_0	925.0	0.486486	0.500088
Угол нашивки, град_90	925.0	0.513514	0.500088
Содержание эпоксидных групп, г	925.0	48.129428	13.667150
Потребление смолы, г/м2 без ЭГ	925.0	168.856094	44.576887

		min	25%
50% \			
Соотношение матрица-наполнитель	0.547391	2.314541	
2.903233			
Плотность, кг/м3	1784.482245	1923.255135	
1977.302956			
модуль упругости, ГПа	2.436909	498.519344	
738.736842			
Количество отвердителя, м.%	38.668500	92.834720	
111.157879			
Поверхностная плотность, г/м2	0.603740	267.140817	
460.721126			
Модуль упругости при растяжении, ГПа	65.793845	71.279418	
73.253725			
Прочность при растяжении, МПа	1250.392802	2149.974687	
2456.395009			
Шаг нашивки	0.037639	5.117477	
6.947322			
Плотность нашивки	28.661632	50.280645	
57.615327			
Угол нашивки, град_0	0.000000	0.000000	
0.000000			
Угол нашивки, град_90	0.000000	0.000000	
1.000000			
Содержание эпоксидных групп, г	13.158372	38.137716	
48.089010			
Потребление смолы, г/м2 без ЭГ	50.427004	139.606267	
168.753112			

	75%	max
Соотношение матрица-наполнитель	3.546415	5.314144
Плотность, кг/м3	2020.828331	2161.565216
модуль упругости, ГПа	956.246864	1628.000000
Количество отвердителя, м.%	130.163998	181.828448
Поверхностная плотность, г/м2	694.589685	1291.340115
Модуль упругости при растяжении, ГПа	75.309657	81.203147
Прочность при растяжении, МПа	2751.235671	3636.892992
Шаг нашивки	8.606411	13.732404
Плотность нашивки	64.854936	86.012427
Угол нашивки, град_0	1.000000	1.000000
Угол нашивки, град_90	1.000000	1.000000
Содержание эпоксидных групп, г	57.358358	84.967714
Потребление смолы, г/м2 без ЭГ	199.907147	284.828567

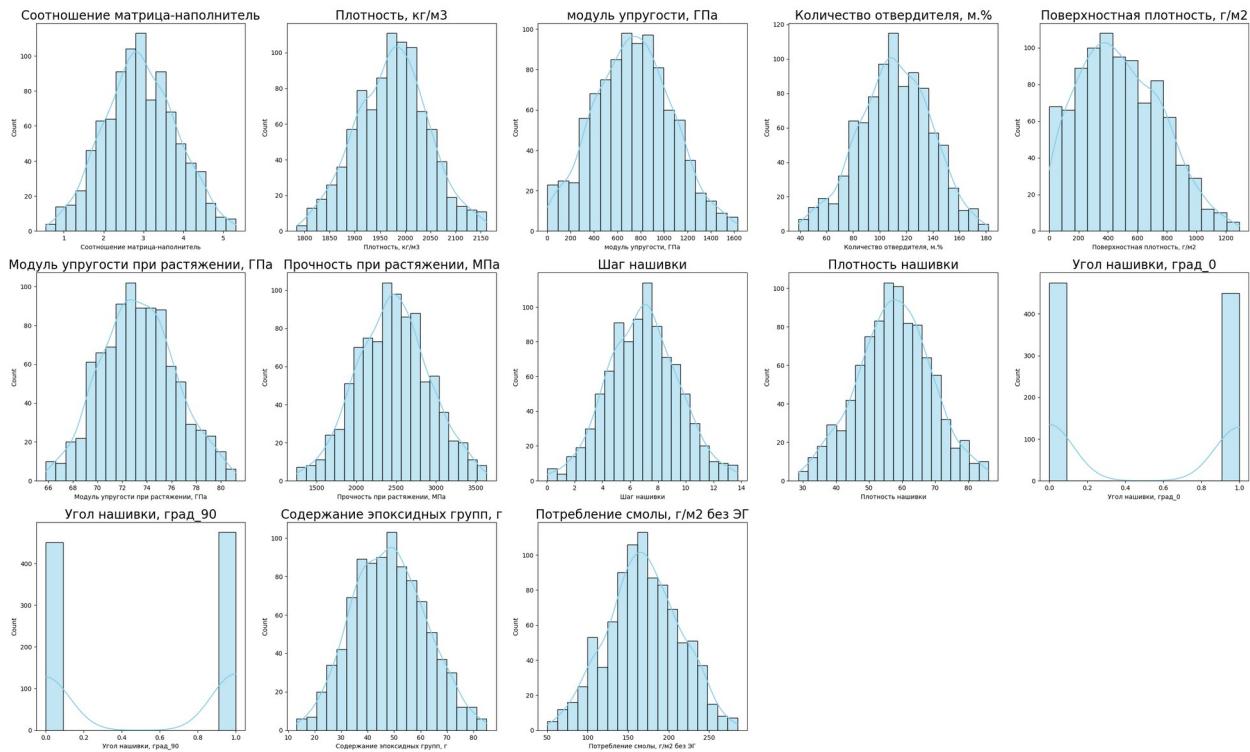
Строим графики гистограмм распределения по каждому столбцу (переменной) в датасете

```
# Строим холст для нанесения 13 графиков (по количеству столбцов).
a = 5 # количество графиков в строке холста
b = 5 # количество график в столбце холста
c = 1 # порядковый номер графика

plt.figure(figsize = (35,35))
plt.suptitle('Гистограммы переменных', fontsize = 25)

for i in df:
    plt.subplot(a, b, c)
    sns.histplot(data=df[i], kde=True, color="skyblue")
    plt.title(i, size = 20)
    c = c + 1
```

## Гистограммы переменных



Т.к. ранее переменная "Угол нашивки, град" кодирована с помощью OneHotEncoder и разнесена на переменные "Угол нашивки, град\_0" и "Угол нашивки, град\_90", то их распределение бимодально, т.е. содержит только два значения - 0 и 1. Отмечаем, что на гистограмме переменной "Поверхностная плотность, г/м<sup>2</sup>" наблюдается асимметрия вправо. По остальным переменным распределение напоминает нормальное.

Проводим статистический тест Шапиро-Уилка, который позволяет проверить гипотезу о нормальности распределения данных. Используем модуль stats.

```

for i in list(df.columns.values):
    stat, p = stats.shapiro(df[i])
    print(i)
    print('Statistics=%.3f, p-value=%.3f' % (stat, p))

    alpha = 0.05
    if p > alpha:
        print('Принять гипотезу о нормальности распределения\n')
    else:
        print('Отклонить гипотезу о нормальности распределения\n')

```

Соотношение матрица-наполнитель  
 Statistics=0.997, p-value=0.098  
 Принять гипотезу о нормальности распределения

Плотность, кг/м<sup>3</sup>  
Statistics=0.997, p-value=0.087  
Принять гипотезу о нормальности распределения

модуль упругости, ГПа  
Statistics=0.995, p-value=0.006  
Отклонить гипотезу о нормальности распределения

Количество отвердителя, м.%  
Statistics=0.997, p-value=0.045  
Отклонить гипотезу о нормальности распределения

Поверхностная плотность, г/м<sup>2</sup>  
Statistics=0.978, p-value=0.000  
Отклонить гипотезу о нормальности распределения

Модуль упругости при растяжении, ГПа  
Statistics=0.996, p-value=0.013  
Отклонить гипотезу о нормальности распределения

Прочность при растяжении, МПа  
Statistics=0.998, p-value=0.172  
Принять гипотезу о нормальности распределения

Шаг нашивки  
Statistics=0.998, p-value=0.312  
Принять гипотезу о нормальности распределения

Плотность нашивки  
Statistics=0.997, p-value=0.066  
Принять гипотезу о нормальности распределения

Угол нашивки, град\_0  
Statistics=0.636, p-value=0.000  
Отклонить гипотезу о нормальности распределения

Угол нашивки, град\_90  
Statistics=0.636, p-value=0.000  
Отклонить гипотезу о нормальности распределения

Содержание эпоксидных групп, г  
Statistics=0.996, p-value=0.015  
Отклонить гипотезу о нормальности распределения

Потребление смолы, г/м<sup>2</sup> без ЭГ  
Statistics=0.996, p-value=0.037  
Отклонить гипотезу о нормальности распределения

На основании гистограмм распределения и статистического теста Шапиро-Уилка делаем выводы о том, что значения переменных "Соотношение матрица-наполнитель", "Плотность, кг/м<sup>3</sup>", "Прочность при растяжении, МПа", "Шаг нашивки", "Плотность нашивки" распределены нормально.

Значения переменных "модуль упругости, ГПа", "Количество отвердителя, м.%", "Модуль упругости при растяжении, ГПа", "Поверхностная плотность, г/м<sup>2</sup>", "Угол нашивки, град\_0", "Угол нашивки, град\_0", "Содержание эпоксидных групп, г", "Потребление смолы, г/м<sup>2</sup> без ЭГ" распределены не нормально.

Если данные не распределены нормально, то применение параметрических статистических тестов, которые предполагают нормальность, может привести к неверным результатам. В таких случаях рекомендуется использовать непараметрические методы, которые не делают предположений о распределении данных.

Необходимо отметить, что после выделения отдельных новых признаков датасета, нормализации и устранения выбросов часть данных перестали быть распределенными нормальным образом. Были предприняты попытки нормализации распределения данных с помощью распределения Бокса-Кокса, логарифмизации и извлечения квадратного корня, но такие попытки не привели к нормализации распределения. В связи с чем принято решения работать с данными в имеющимся виде.

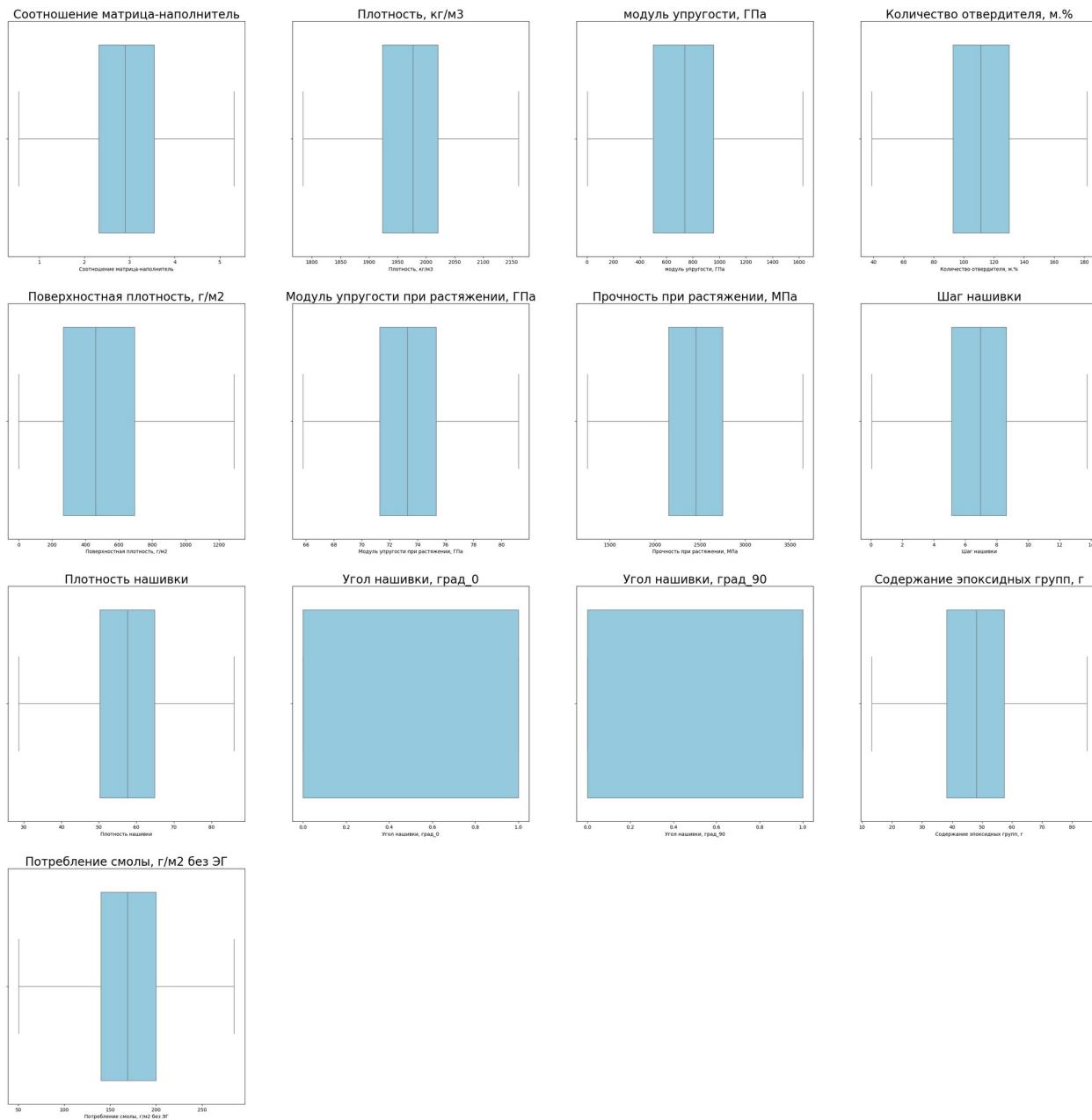
Строим диаграммы размаха по каждой переменной в итоговом датасете

```
# Строим холст для нанесения 13 графиков (по количеству столбцов).
a = 4 # количество графиков в строке холста
b = 4 # количество график в столбце холста
c = 1 # порядковый номер графика

plt.figure(figsize = (40,40))
plt.suptitle('Диаграммы размаха', fontsize = 30)

for i in df:
    plt.subplot(a, b, c)
    sns.boxplot(data = df[i], color = "skyblue", orient='h')
    plt.title(i, size = 24)
    c = c + 1
```

### Диаграммы размаха

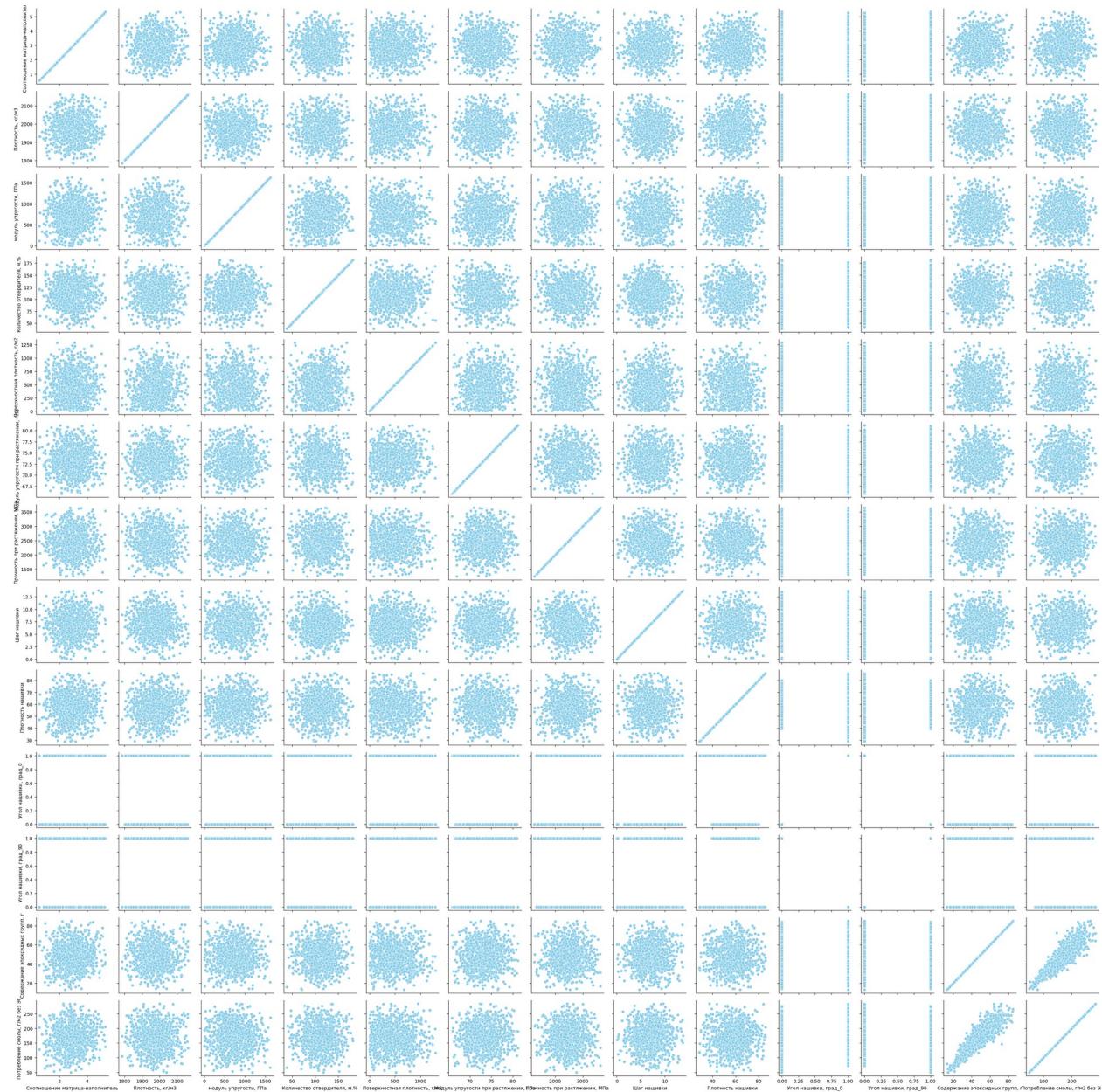


Делаем вывод об отсутствии выбросов на основании диаграмм размаха.

Строим попарные диаграммы рассеяния каждой переменной в датасете

```
parn_gr_rass = sns.PairGrid(df[df.columns])
parn_gr_rass.map(sns.scatterplot, color="skyblue")
plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```



Обнаруживаем сильную корреляцию между "Потреблением смолы, г/м<sup>2</sup>" и "Содержанием эпоксидных групп, г" с коэффициентом корреляции 0.86 (очень сильная связь), что не удивительно т.к. на основании информации о связи эпоксидных групп и их содержанием в смолах, данные признаки были доработаны.

По остальным признакам обнаружено отсутствие как положительной, так и отрицательной корреляции между переменными. Линейная связь обнаружена только между переменными "Потреблением смолы, г/м<sup>2</sup>" и "Содержанием эпоксидных групп, г", что может повлиять на исследования связанные с выявлением линейный связей.

Дополнительно с помощью метода pd.DataFrame.corr() вычисляем попарную корреляцию переменных в датафрейме.

```
df.corr()
```

Соотношение матрица-наполнитель	
\	1.000000
Соотношение матрица-наполнитель	1.000000
Плотность, кг/м3	0.006542
модуль упругости, ГПа	0.055564
Количество отвердителя, м.%	-0.003040
Поверхностная плотность, г/м2	0.011418
Модуль упругости при растяжении, ГПа	-0.025333
Прочность при растяжении, МПа	0.017655
Шаг нашивки	0.042252
Плотность нашивки	0.044812
Угол нашивки, град_0	0.027447
Угол нашивки, град_90	-0.027447
Содержание эпоксидных групп, г	0.066277
Потребление смолы, г/м2 без ЭГ	0.069049

Плотность, кг/м3			модуль
упругости, ГПа \			
Соотношение матрица-наполнитель	0.006542		
0.055564			
Плотность, кг/м3	1.000000		
0.000332			
модуль упругости, ГПа	0.000332		
1.000000			
Количество отвердителя, м.%	-0.045209		
0.048065			
Поверхностная плотность, г/м2	0.064960		-
0.010099			
Модуль упругости при растяжении, ГПа	-0.019442		
0.020775			
Прочность при растяжении, МПа	-0.088375		
0.027288			
Шаг нашивки	-0.049241		

0.009450	
Плотность нашивки	0.078473
0.076889	
Угол нашивки, град_0	0.054210
0.033468	
Угол нашивки, град_90	-0.054210
0.033468	
Содержание эпоксидных групп, г	-0.025589
0.007213	
Потребление смолы, г/м2 без ЭГ	-0.025332
0.002584	

	Количество отвердителя, м.% \
Соотношение матрица-наполнитель	-0.003040
Плотность, кг/м3	-0.045209
модуль упругости, ГПа	0.048065
Количество отвердителя, м.%	1.000000
Поверхностная плотность, г/м2	0.039558
Модуль упругости при растяжении, ГПа	-0.053024
Прочность при растяжении, МПа	-0.068661
Шаг нашивки	-0.013678
Плотность нашивки	0.006798
Угол нашивки, град_0	-0.037753
Угол нашивки, град_90	0.037753
Содержание эпоксидных групп, г	-0.005353
Потребление смолы, г/м2 без ЭГ	-0.008743

	Поверхностная плотность, г/м2 \
Соотношение матрица-наполнитель	0.011418
Плотность, кг/м3	0.064960
модуль упругости, ГПа	-0.010099
Количество отвердителя, м.%	0.039558
Поверхностная плотность, г/м2	1.000000
Модуль упругости при растяжении, ГПа	0.038745
Прочность при растяжении, МПа	-0.016811
Шаг нашивки	0.039422
Плотность нашивки	-0.031263
Угол нашивки, град_0	-0.049201
Угол нашивки, град_90	0.049201
Содержание эпоксидных групп, г	-0.015456
Потребление смолы, г/м2 без ЭГ	-0.006632

	Модуль упругости при растяжении, ГПа \
Соотношение матрица-наполнитель	-
0.025333	
Плотность, кг/м3	-
0.019442	
модуль упругости, ГПа	-
0.020775	

Количество отвердителя, м.%	-
0.053024	
Поверхностная плотность, г/м <sup>2</sup>	
0.038745	
Модуль упругости при растяжении, ГПа	
1.000000	
Прочность при растяжении, МПа	-
0.006631	
Шаг нашивки	-
0.005490	
Плотность нашивки	
0.005693	
Угол нашивки, град_0	-
0.035956	
Угол нашивки, град_90	
0.035956	
Содержание эпоксидных групп, г	
0.058409	
Потребление смолы, г/м <sup>2</sup> без ЭГ	
0.045717	

	Прочность при растяжении, МПа \
Соотношение матрица-наполнитель	0.017655
Плотность, кг/м <sup>3</sup>	-0.088375
модуль упругости, ГПа	0.027288
Количество отвердителя, м.%	-0.068661
Поверхностная плотность, г/м <sup>2</sup>	-0.016811
Модуль упругости при растяжении, ГПа	-0.006631
Прочность при растяжении, МПа	1.000000
Шаг нашивки	-0.049983
Плотность нашивки	0.014727
Угол нашивки, град_0	-0.026231
Угол нашивки, град_90	0.026231
Содержание эпоксидных групп, г	0.001291
Потребление смолы, г/м <sup>2</sup> без ЭГ	0.008875

	Шаг нашивки	Плотность
нашивки \		
Соотношение матрица-наполнитель	0.042252	0.044812
Плотность, кг/м <sup>3</sup>	-0.049241	0.078473
модуль упругости, ГПа	0.009450	0.076889
Количество отвердителя, м.%	-0.013678	0.006798
Поверхностная плотность, г/м <sup>2</sup>	0.039422	-0.031263
Модуль упругости при растяжении, ГПа	-0.005490	0.005693

Прочность при растяжении, МПа	-0.049983	0.014727
Шаг нашивки	1.000000	0.002886
Плотность нашивки	0.002886	1.000000
Угол нашивки, град_0	-0.026675	-0.084994
Угол нашивки, град_90	0.026675	0.084994
Содержание эпоксидных групп, г	0.014175	-0.005769
Потребление смолы, г/м2 без ЭГ	0.007792	0.007016
Угол нашивки, град_0 \		
Соотношение матрица-наполнитель	0.027447	
Плотность, кг/м3	0.054210	
модуль упругости, ГПа	0.033468	
Количество отвердителя, м.%	-0.037753	
Поверхностная плотность, г/м2	-0.049201	
Модуль упругости при растяжении, ГПа	-0.035956	
Прочность при растяжении, МПа	-0.026231	
Шаг нашивки	-0.026675	
Плотность нашивки	-0.084994	
Угол нашивки, град_0	1.000000	
Угол нашивки, град_90	-1.000000	
Содержание эпоксидных групп, г	-0.022410	
Потребление смолы, г/м2 без ЭГ	-0.004355	
Угол нашивки, град_90 \		
Соотношение матрица-наполнитель	-0.027447	
Плотность, кг/м3	-0.054210	
модуль упругости, ГПа	-0.033468	
Количество отвердителя, м.%	0.037753	
Поверхностная плотность, г/м2	0.049201	
Модуль упругости при растяжении, ГПа	0.035956	
Прочность при растяжении, МПа	0.026231	
Шаг нашивки	0.026675	
Плотность нашивки	0.084994	
Угол нашивки, град_0	-1.000000	
Угол нашивки, град_90	1.000000	
Содержание эпоксидных групп, г	0.022410	
Потребление смолы, г/м2 без ЭГ	0.004355	
Содержание эпоксидных групп,		
г \		
Соотношение матрица-наполнитель	0.066277	
Плотность, кг/м3	-0.025589	

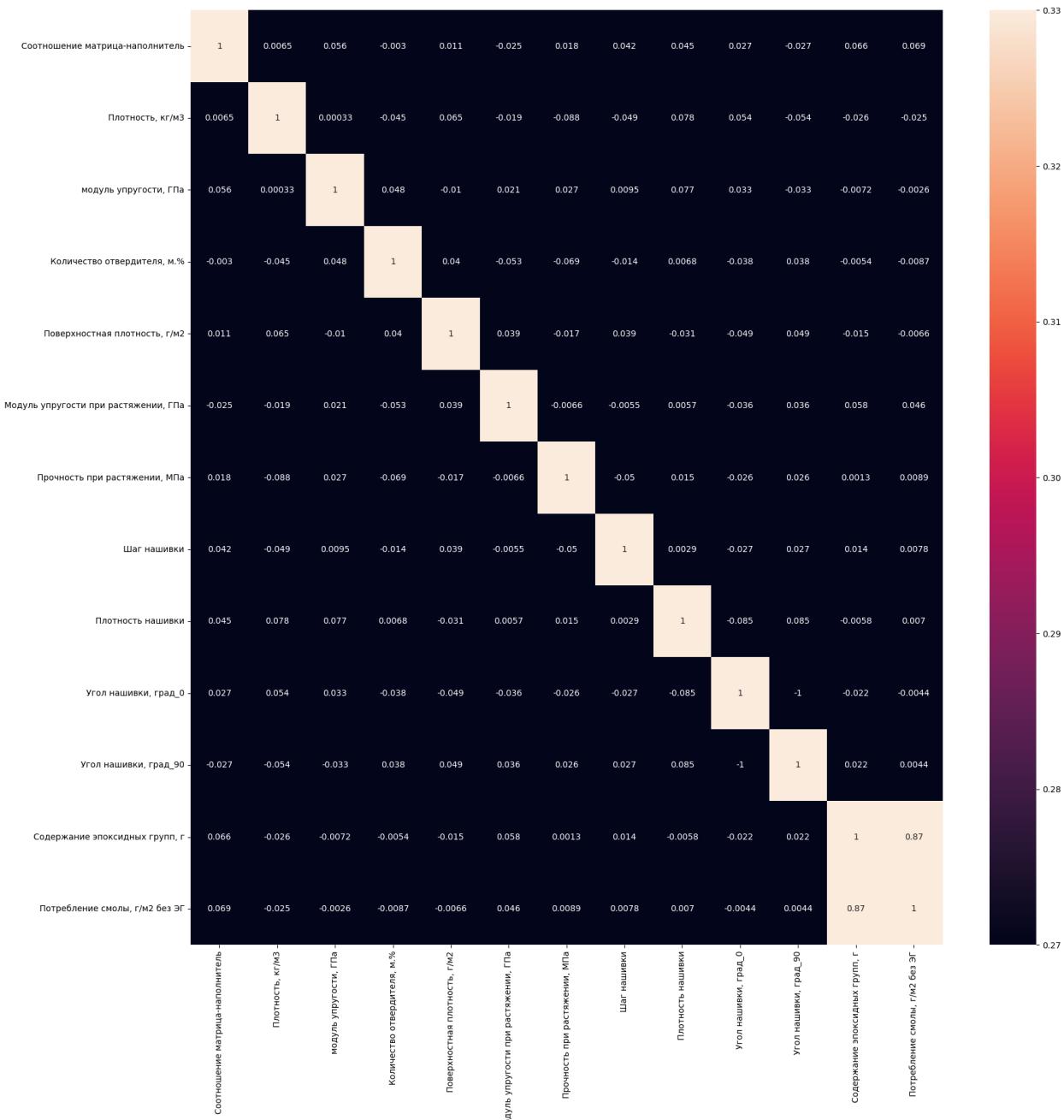
модуль упругости, ГПа	-0.007213
Количество отвердителя, м.%	-0.005353
Поверхностная плотность, г/м2	-0.015456
Модуль упругости при растяжении, ГПа	0.058409
Прочность при растяжении, МПа	0.001291
Шаг нашивки	0.014175
Плотность нашивки	-0.005769
Угол нашивки, град_0	-0.022410
Угол нашивки, град_90	0.022410
Содержание эпоксидных групп, г	1.000000
Потребление смолы, г/м2 без ЭГ	0.866718

	Потребление смолы, г/м2 без ЭГ
Соотношение матрица-наполнитель	0.069049
Плотность, кг/м3	-0.025332
модуль упругости, ГПа	-0.002584
Количество отвердителя, м.%	-0.008743
Поверхностная плотность, г/м2	-0.006632
Модуль упругости при растяжении, ГПа	0.045717
Прочность при растяжении, МПа	0.008875
Шаг нашивки	0.007792
Плотность нашивки	0.007016
Угол нашивки, град_0	-0.004355
Угол нашивки, град_90	0.004355
Содержание эпоксидных групп, г	0.866718
Потребление смолы, г/м2 без ЭГ	1.000000

Дополнительно отобразим тепловую матрицу корреляции переменных датасета, т.к. визуализация данных позволяет лучше обнаружить наличие корреляции между переменными.

```
#Отобразим тепловую матрицу для наглядности
fig = plt.figure(figsize = (20,20))
sns.heatmap(df.corr(), annot = True, vmax=0.3, vmin=-0.3)

<Axes: >
```



Обнаруживаем сильную корреляцию между "Потреблением смолы, г/м<sup>2</sup>" и "Содержанием эпоксидных групп, г" с коэффициентом корреляции 0.87 (очень сильная связь), что не удивительно т.к. на основании информации о связи эпоксидных групп и их содержанием в смолах, данные признаки были доработаны. В отношении иных переменных значение коэффициента корреляции близко к 0, что указывает на отсутствие корреляции.

Проведен разведывательный анализ и предобратока датасета. Добавлено описание столбцов и их описание. Проведена работа с анализом, исключением и модификацией переменных:

- Исключена переменная "Температура вспышки, С\_2", как не оказывающая влияние на предсказание значений "Модуль упругости при растяжении, ГПа", "Прочность при растяжении, МПа", "Соотношение матрица-наполнитель".
- Переменная "Угол нашивки, град" переведена в раздел категориальных переменных и кодирована с помощью OneHotEncoder в переменные "Угол нашивки, град\_0" и "Угол нашивки, град\_90".
- Так же проведен анализ взаимосвязи переменных "Содержание эпоксидных групп, %\_2" и "Потребление смолы, г/м2", в результате чего переменные были изменены в переменные "Содержание эпоксидных групп, г" и "Потребление смолы, г/м2 без ЭГ". Так же выдвинуто предположение о наличии связей между плотностью, отвердителем, эпоксидными группами и смолами - в данной части необходима консультация с экспертом. Проведена визуализация данных. Построены гистограммы распределения данных с KDE, диаграммы размаха, диаграммы рассеяния, тепловые карты. Проведен статистический тест для подтверждения нормальности распределения. Исключены выбросы. Исследована корреляция переменных. Проведена нормализация данных.

Необходимо отметить, что многие переменные не распределены нормально, практически полное отсутствие корреляции между переменными и линейной связи, что может ухудшить предсказательную способность прогнозирующих моделей машинного обучения.

Сохраняем итоговый датасет в отдельный файл.

```
df.to_excel('db_itog_bez_norm.xlsx')
```