

**МИНОБРНАУКИ РОССИИ**  
**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ**  
**ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**  
**«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**  
**Кафедра ИС**

**ОТЧЁТ О ПРОВЕДЕНИИ ИССЛЕДОВАНИЯ**

**по дисциплине «Интеллектуальный анализ данных»**

**ТЕМА:** зависимость между наличием диабета и медицинских показателей в  
крови

Студент гр. 0374

Подтергера А.А.

Преподаватель

Татчина Я.А.

Санкт-Петербург

2025

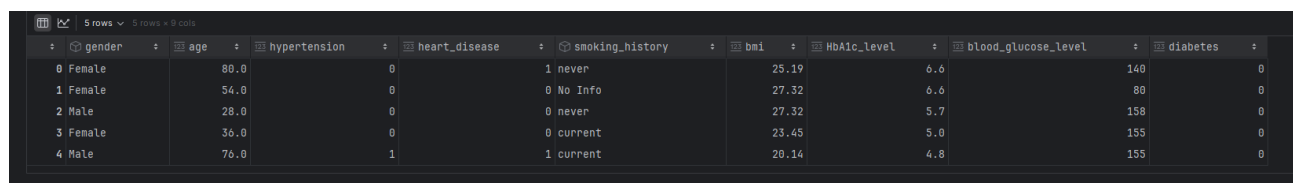
Целью исследования является выявление взаимосвязей и зависимостей, следовательно будет применён описательный анализ данных.

Будет проведено исследование наличия диабета у человека от значений разных показателей в крови и наличием вредных привычек. Набор данных, заимствованных с интернет-ресурса <https://www.kaggle.com>

Данные имеют следующий тип:

Название столбца	Тип данных	Описание
<b>gender</b>	object (строка)	Пол пациента: Male, Female, Other.
<b>age</b>	float	Возраст пациента в годах.
<b>hypertension</b>	int (0/1)	Наличие гипертонии: 0 — нет, 1 — да.
<b>heart_disease</b>	int (0/1)	Наличие сердечного заболевания: 0 — нет, 1 — да.
<b>smoking_history</b>	object (строка)	История курения: - never — не курил - former — курил раньше - current — курит сейчас - ever — когда-либо курил - not current — не курит сейчас - No Info — нет информации
<b>bmi</b>	float	Индекс массы тела (Body Mass Index), рассчитывается как вес (кг) / рост <sup>2</sup> (м <sup>2</sup> ). Нормой считается 18.5–24.9.
<b>HbA1c_level</b>	float	Уровень гликированного гемоглобина (HbA1c), обычно в диапазоне 4.0–14.0. Показатель среднего уровня сахара в крови за 2–3 месяца.
<b>blood_glucose_level</b>	float или int	Уровень глюкозы в крови в мг/дл. Нормой считается около 70–140 мг/дл.
<b>diabetes</b>	int (0/1)	Целевой признак: наличие диабета. 1 — диабет есть, 0 — нет.

Рассмотрим набор данных с количеством записей в 100 000:



gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0 Female	80.0	0	0	1 never	25.19	6.6	140	0
1 Female	54.0	0	0	0 No Info	27.32	6.6	89	0
2 Male	28.0	0	0	0 never	27.32	5.7	158	0
3 Female	36.0	0	0	0 current	23.45	5.0	155	0
4 Male	76.0	1	1	1 current	20.14	4.8	155	0

Рисунок 1 – Первые 5 записей

1 df.dtypes # определяем тип данных столбцов df

✓ [66] 18ms

9 rows 9 rows x 1 cols

	<unnamed>
gender	object
age	float64
hypertension	int64
heart_disease	int64
smoking_history	object
bmi	float64
HbA1c_level	float64
blood_glucose_level	int64
diabetes	int64

1 df.isnull().sum() # проверяем на нулевые значения

✓ [67] 17ms

9 rows 9 rows x 1 cols

	<unnamed>
gender	0
age	0
hypertension	0
heart_disease	0
smoking_history	0
bmi	0
HbA1c_level	0
blood_glucose_level	0
diabetes	0

1 df.nunique() # уникальные значения

✓ [51] 80ms

9 rows 9 rows x 1 cols

	<unnamed>
gender	3
age	102
hypertension	2
heart_disease	2
smoking_history	6
bmi	4247
HbA1c_level	18
blood_glucose_level	18
diabetes	2

Рисунок 2 – Первичный анализ

Первичный анализ данных показал, что все признаки в датасете заполнены — пропусков нет. Типы данных соответствуют содержимому: числовые признаки представлены числами с плавающей точкой или целыми, категориальные — строками.

Категориальные признаки пол (gender) и историю курения (smoking\_history), каждый из которых содержит несколько уникальных значений. Числовые признаки, такие как возраст, индекс массы тела, уровень гликированного гемоглобина и уровень глюкозы в крови, имеют широкий

диапазон значений и пригодны для анализа. Целевой признак — наличие диабета (diabetes)(да/нет).

Датасет готов для дальнейшего анализа: визуализации, выявления закономерностей.

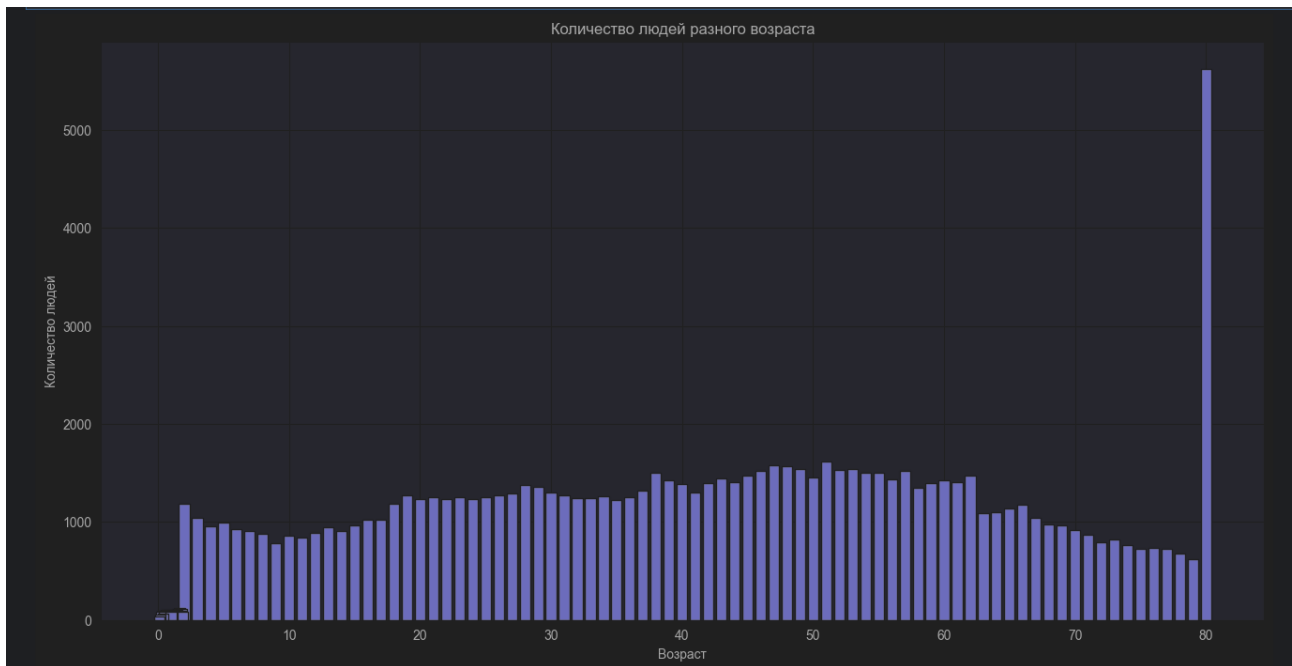


Рисунок 3 – Количество людей разного возраста

В датасете распределение людей по возрасту **неравномерное**. Больше всего наблюдений приходится на взрослых и пожилых людей — основную часть выборки составляют пациенты в возрасте от примерно 30 до 70 лет.

Это значит, что данные **смещены в сторону пожилого возраста**, что логично, поскольку риск диабета и сопутствующих заболеваний возрастает с возрастом.

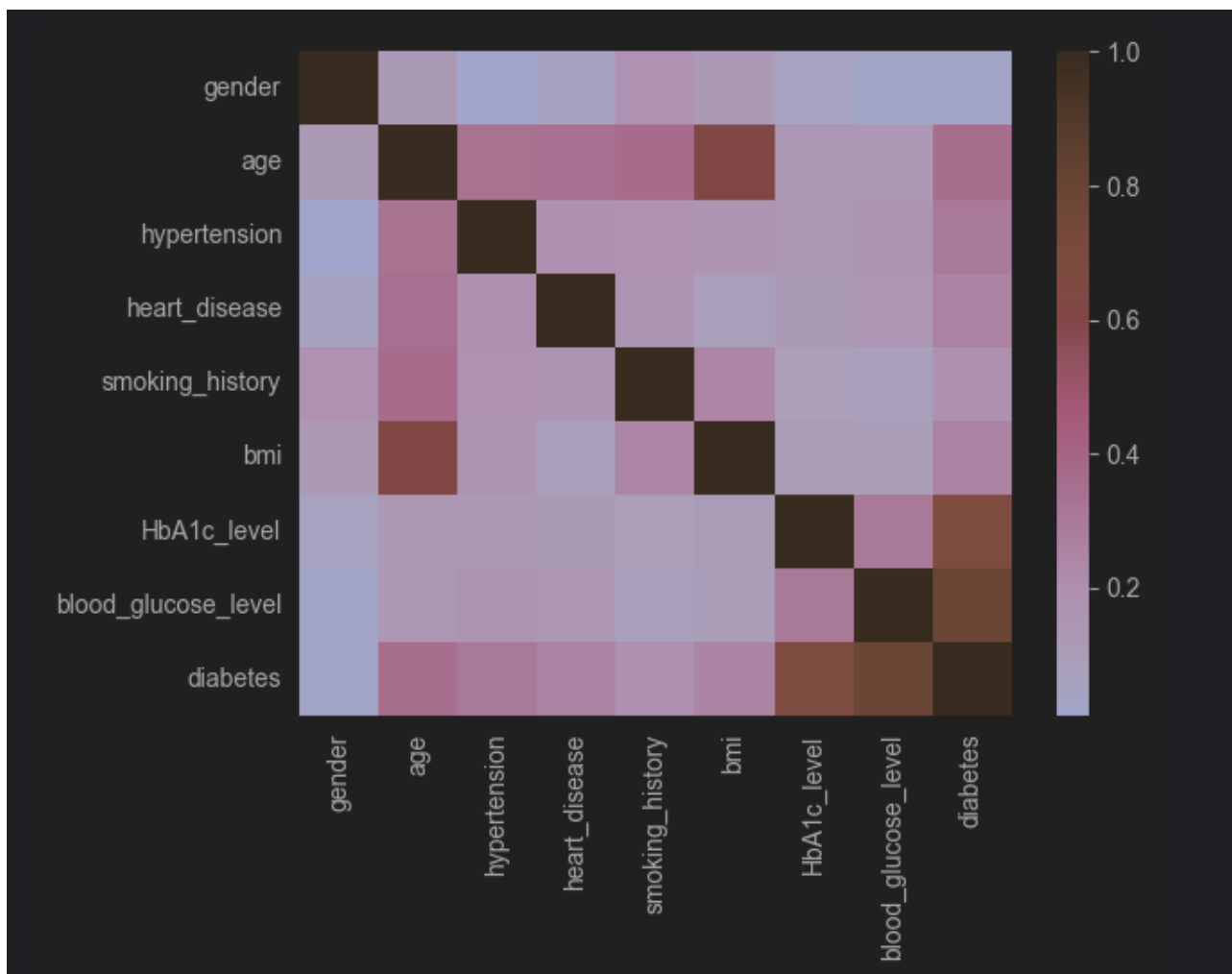


Рисунок 4 – Тепловая карта

**Наибольшая корреляция с диабетом** наблюдается у признаков:

- **blood\_glucose\_level** — положительная корреляция: чем выше уровень сахара в крови, тем выше вероятность наличия диабета.
- **HbA1c\_level** — также положительная связь: повышенный уровень гликированного гемоглобина характерен для диабетиков.
- **age** — умеренно положительная корреляция: риск диабета возрастает с возрастом.

**Слабая или отсутствующая корреляция** у признаков:

- **gender, smoking\_history, bmi** — слабо связаны напрямую с диабетом, но могут иметь опосредованное влияние в комбинации с другими признаками.

- **hypertension** и **heart\_disease** имеют слабую положительную корреляцию с диабетом.

Далее выполним нормализацию значений по курению. Заменяем значения курящих на 1, в ином случае 0, 2 - нет информации.

```
f_m_alco = pd.crosstab(df['gender'], df['smoking_history'])
f_alco = f_m_alco.iloc[0][1]/(f_m_alco.iloc[0][1]+f_m_alco.iloc[0][0])
m_alco = f_m_alco.iloc[1][1]/(f_m_alco.iloc[1][1]+f_m_alco.iloc[1][0])
print(m_alco)      m_alco
print(f_alco) # процент курящих женщин и процент курящих мужчин, больше курят мужчины
✓ [58] 76ms

0.4174960505529226
0.31066611757438484
```

Рисунок 5 – Процент курящих мужчин и женщин

Мужчины, согласно анализу курят больше.

Построим диаграмму для визуализации:

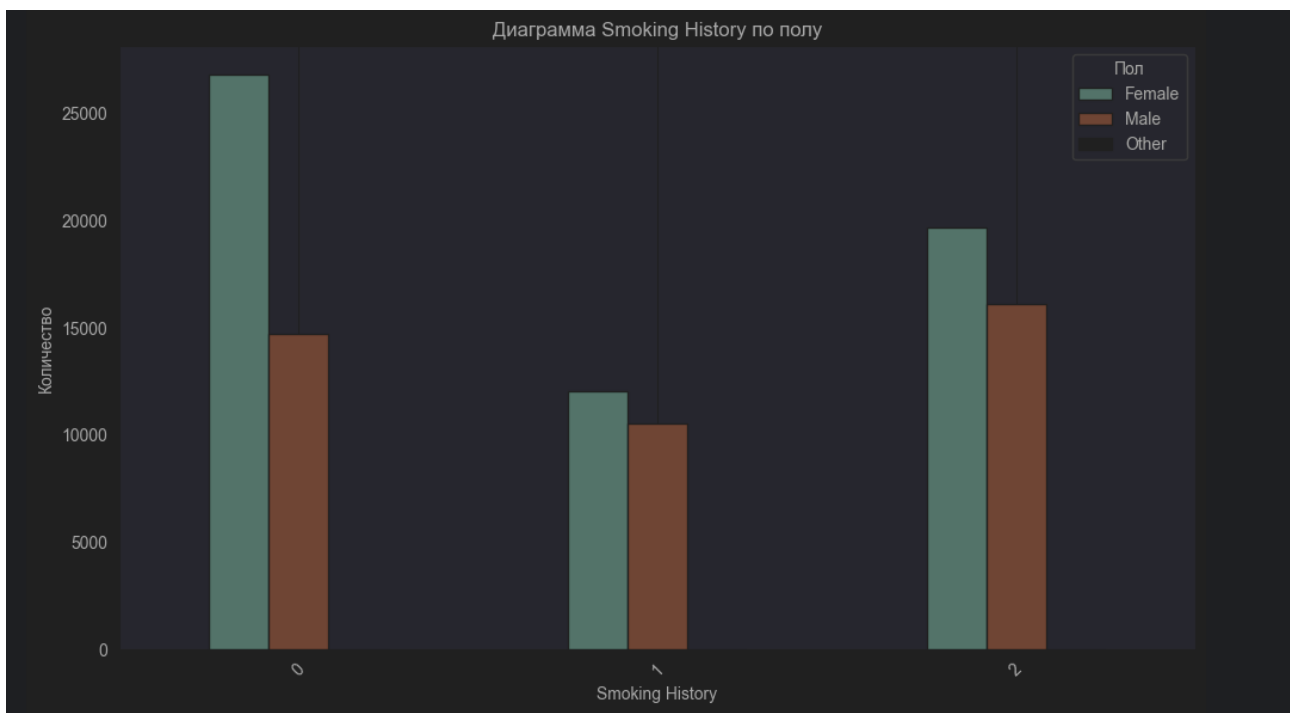


Рисунок 6 – Процент курящих мужчин и женщин

Столбца Other не видно из-за очень низкого количества людей этой группы.

```

1 print(df[df['smoking_history'] == 0]['age'].mean())
2 print(df[df['smoking_history'] == 1]['age'].mean())
3 print(df[df['smoking_history'] == 2]['age'].mean())
4 # средний возраст курящих и некурящих
5 # 0 - не курят, 1 - курят или курили, 2 - нет инфо
✓ [81] 37ms

44.480930142987816
50.329497394223125
33.538036631672995

```

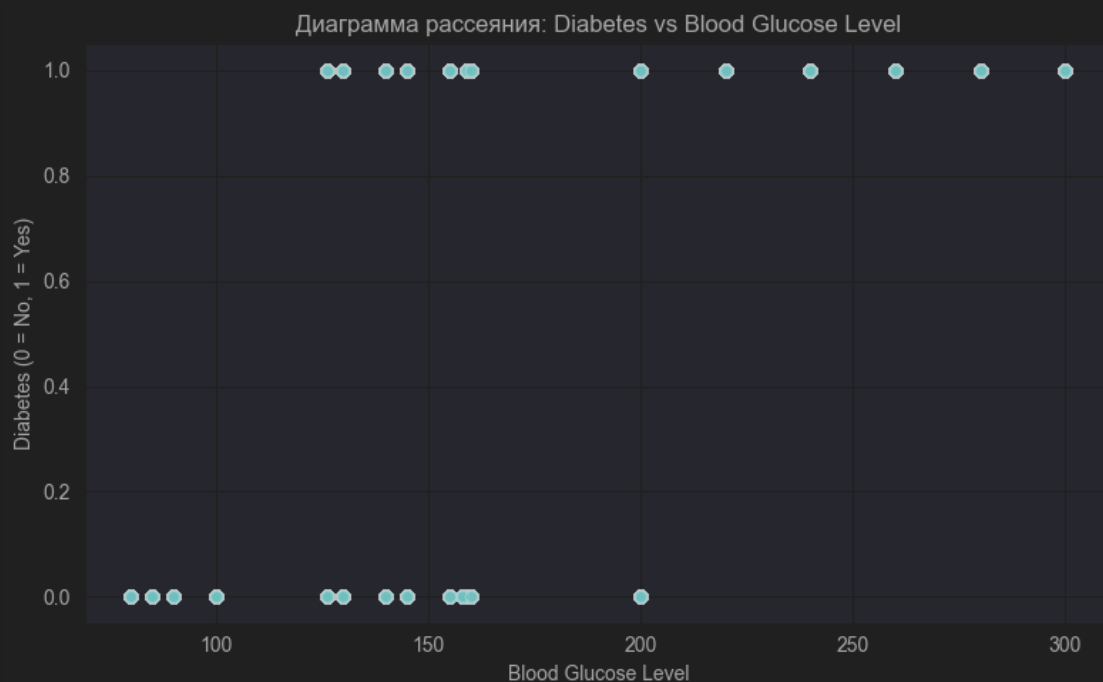
Рисунок 7 – Средний возраст курящих и некурящих

Вычислим корреляцию и построим диаграмму рассеяния. Выводы совпадают с выводами, сделанными по тепловой карте.

```

Корреляция между 'diabetes' и 'blood_glucose_level': 0.4196
Корреляция между 'diabetes' и 'HbA1c_level': 0.4007
Корреляция между 'diabetes' и 'bmi': 0.2144
Корреляция между 'diabetes' и 'age': 0.2580

```



Наблюдаем положительную корреляцию по всем параметрам

Рисунок 8 – Корреляция

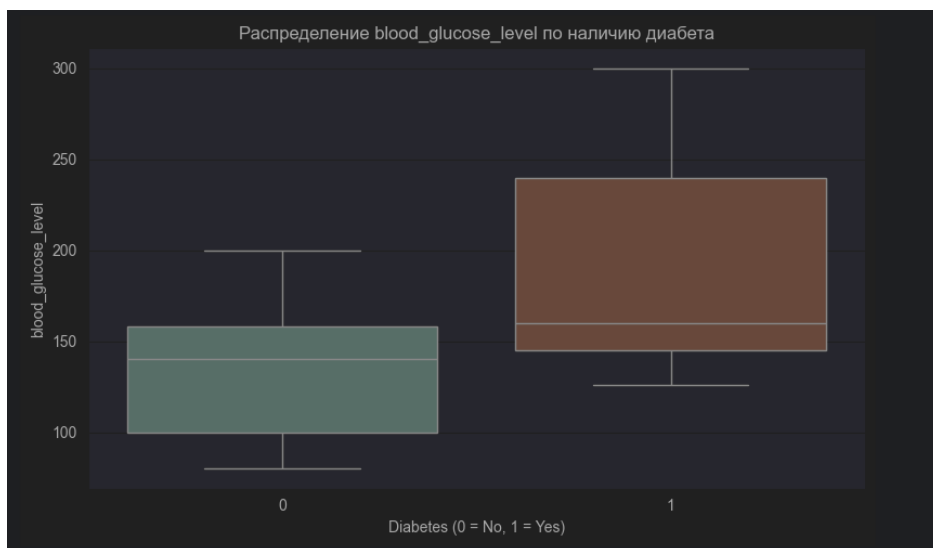


Рисунок 9 – Распределение по *blood\_glucose\_level*

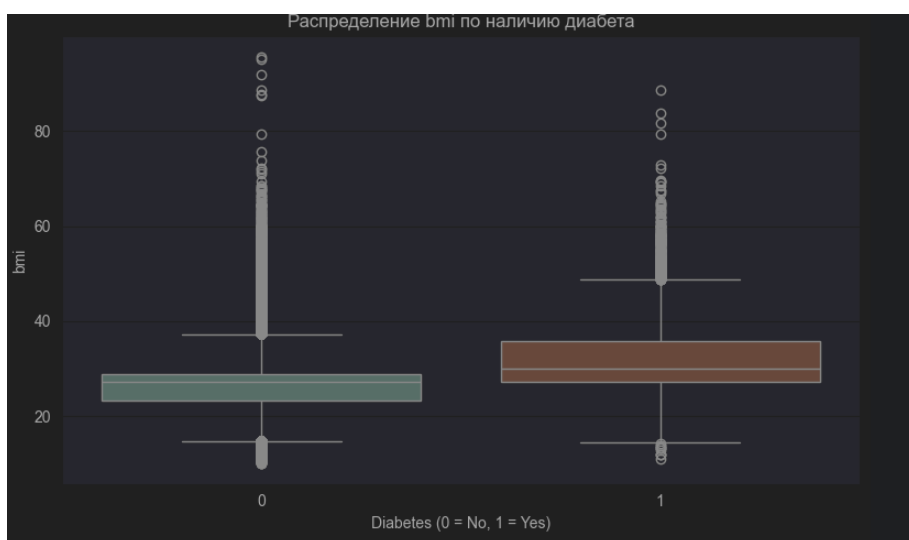


Рисунок 10 – Распределение по *bmi*

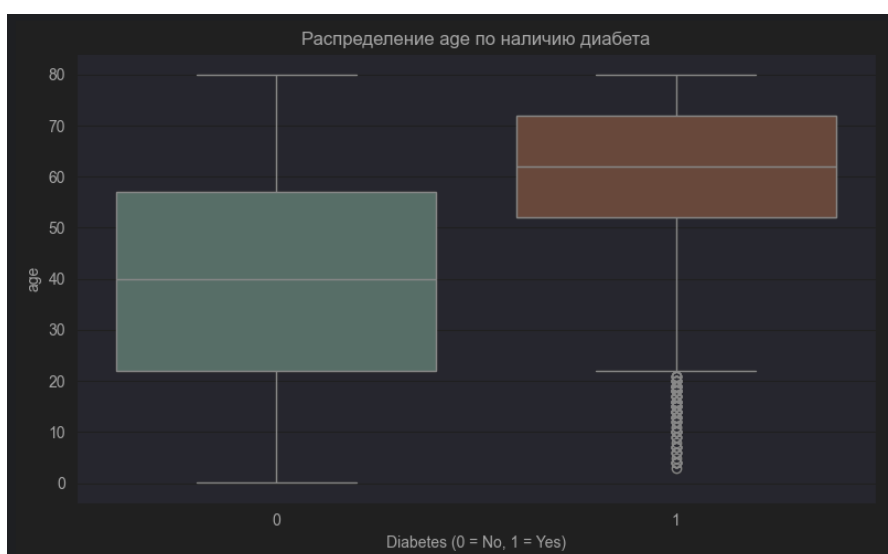


Рисунок 11 – Распределение по *age*



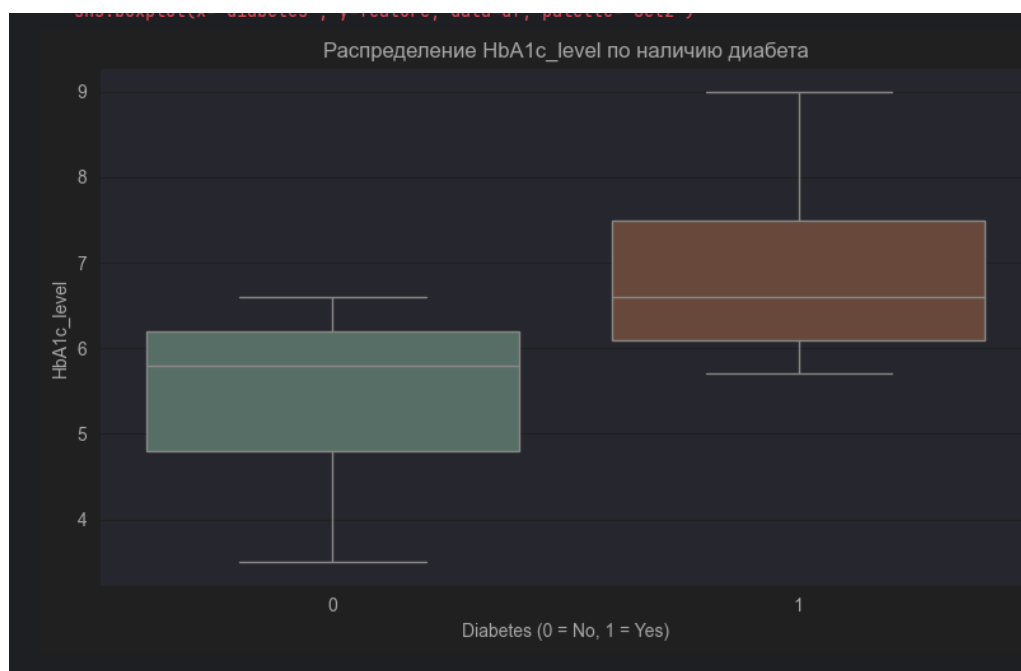


Рисунок 12 – Распределение по HbA1c\_level

### 1. age (возраст)

- **Медиана:** находится в районе среднего возраста (около 45–50 лет).
- **Ящик:** показывает, что основная часть наблюдений (50%) находится между примерно 30 и 60 годами.
- **Усы:** охватывают молодых (около 20 лет) и пожилых (до 80+).
- **Выбросы:** незначительные, в основном в пожилом возрасте.

Вывод: возраст распределён с перекосом к взрослым и пожилым людям.

### 2. bmi (индекс массы тела)

- **Медиана:** немного выше нормы (в районе 27–30).
- **Ящик:** большая часть значений — избыточный вес или лёгкое ожирение.
- **Выбросы:** есть пациенты с экстремально высоким ИМТ (>50), что отражает ожирение высокой степени.

Вывод: большинство имеют ИМТ выше нормы, что повышает риск диабета.

### 3. HbA1c\_level (гликированный гемоглобин)

- **Медиана:** ближе к норме (около 5.5–6.0).
- **Ящик :** указывает на нормальные и пограничные уровни у большинства людей.
- **Выбросы:** высокие значения (7.0–10.0) указывают на диабет.

Вывод: основная масса пациентов с нормальным или чуть повышенным уровнем.

#### 4. blood\_glucose\_level (уровень глюкозы в крови)

- **Медиана:** в верхней части нормы или выше.
- **Ящик:** значения широко варьируются, много пациентов с повышенным уровнем.
- **Выбросы:** выраженные — особенно при уровнях сахара  $>200$ .

Построим диаграмму наличия сердечных заболеваний у людей с диабетом и высоким уровнем сахара.

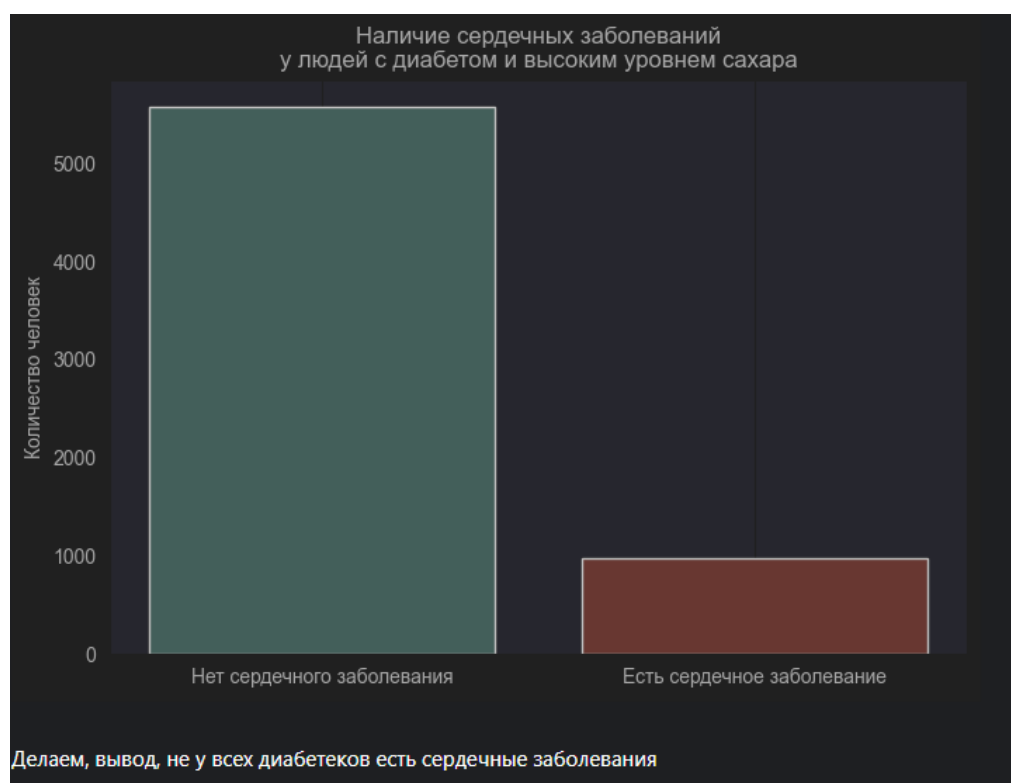


Рисунок 13 – Наличие сердечных заболеваний у людей с диабетом и высоким уровнем сахара

Делаем, вывод, не у всех людей с диабетом есть сердечные заболевания.