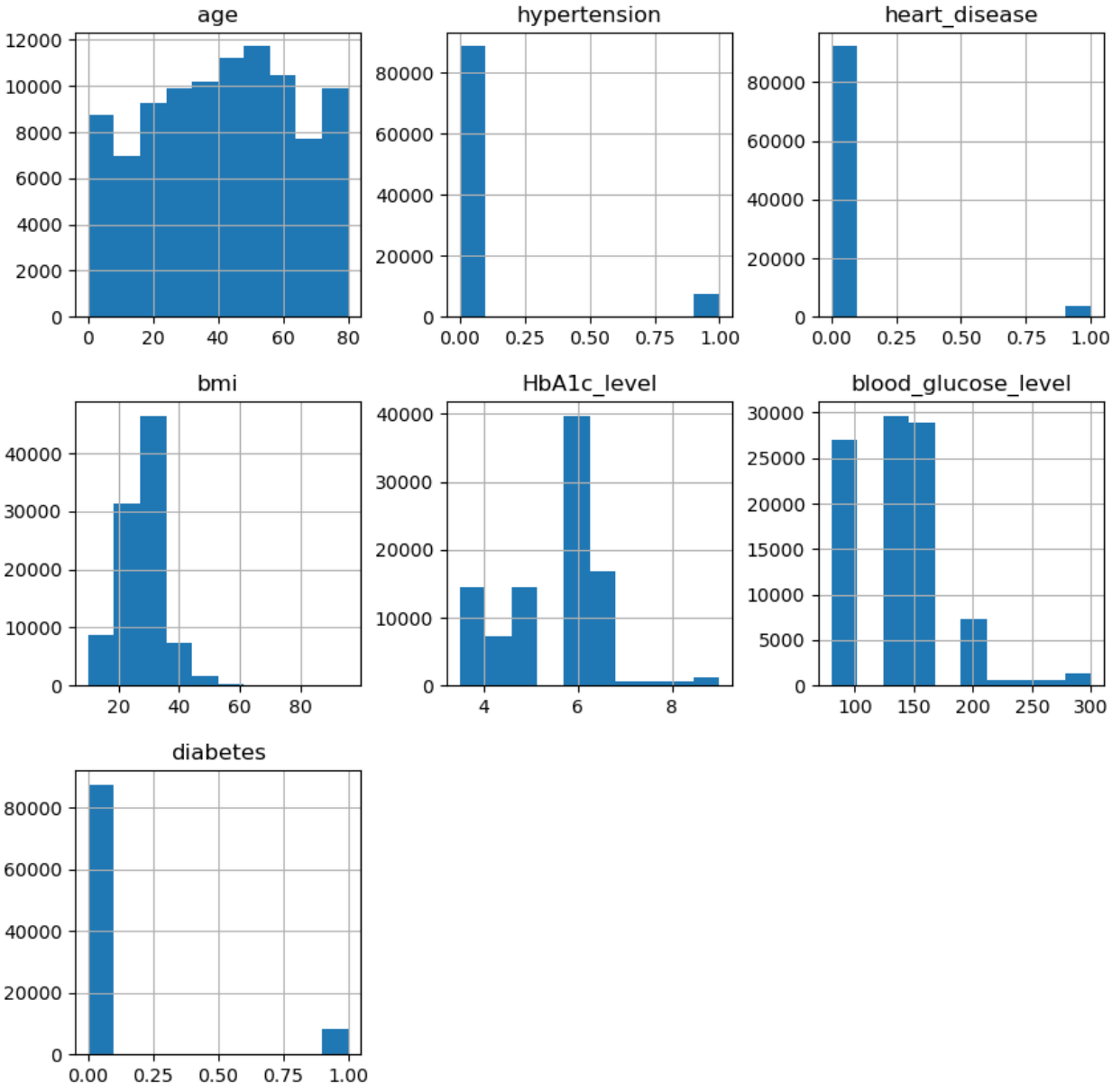
In [1]:

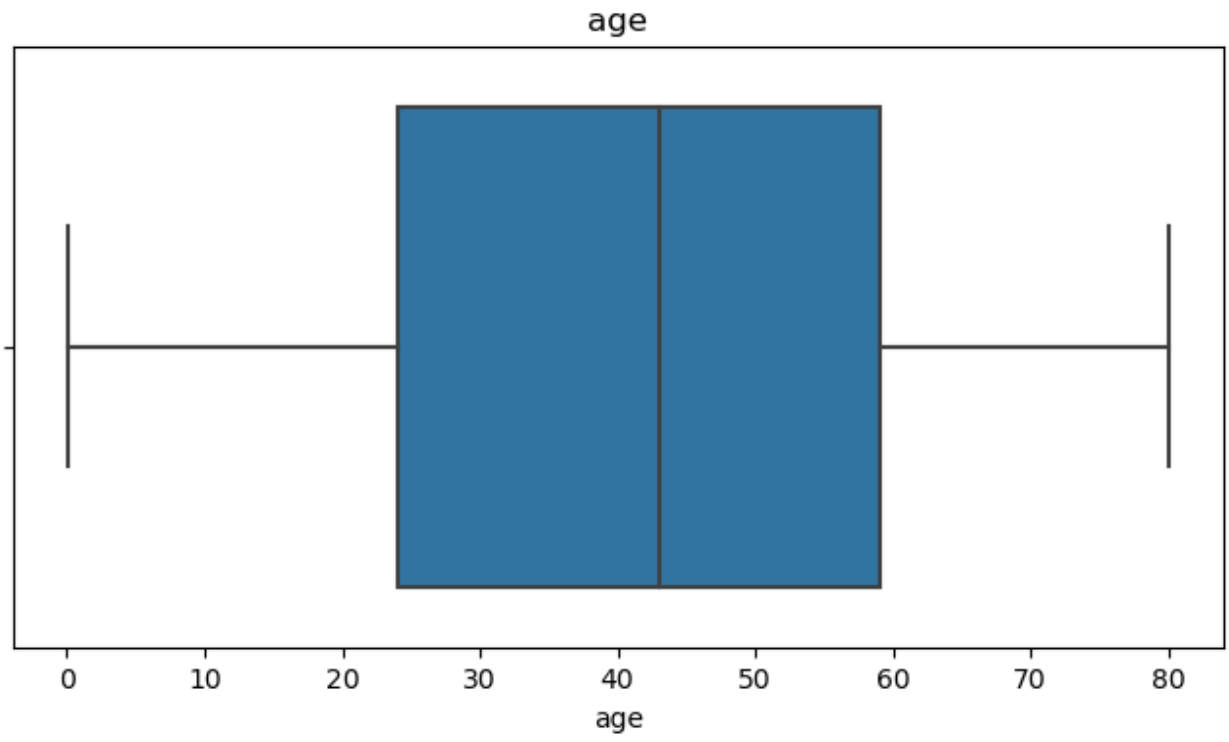
```python
import os
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Загрузка данных
current_dir = os.getcwd()
file_path = os.path.join(current_dir, 'diabetes_prediction_dataset.csv')
df = pd.read_csv(file_path)

# 1. Описательный анализ данных
print("\nОписательный анализ данных:")
print(df.describe())
print(df.info())

# 2. Предобработка данных: Удаление дубликатов
print("\nПредобработка данных: Удаление дубликатов")
initial_rows = df.shape[0]
df.drop_duplicates(inplace=True)
print(f"Удалено дубликатов: {initial_rows - df.shape[0]}")

# 3. Предобработка данных: Отсутствующие значения
print("\nОбработка отсутствующих значений:")
print(f"Количество NaN до:\n{df.isna().sum()}")
df.fillna(df.median(numeric_only=True), inplace=True)
for column in df.select_dtypes(include=['object']).columns:
    df[column].fillna(df[column].mode()[0], inplace=True)
print(f"Количество NaN после:\n{df.isna().sum()}")

# 4. Изменение типа данных
df['age'] = df['age'].astype(float)
df['HbA1c_level'] = df['HbA1c_level'].astype(float)

# Визуализация данных
# Гистограммы числовых переменных
numerical_columns = df.select_dtypes(include=['float64', 'int64']).column
df[numerical_columns].hist(figsize=(10, 10))
plt.show()

# Диаграммы размаха
for column in numerical_columns:
    plt.figure(figsize=(8, 4))
    sns.boxplot(x=df[column])
    plt.title(column)
    plt.show()

# Столбчатые диаграммы для категориальных переменных
categorical_columns = df.select_dtypes(include=['object']).columns
for column in categorical_columns:
    plt.figure(figsize=(8, 4))
    sns.countplot(x=df[column])
    plt.title(column)
    plt.show()

# Сравнение выборок: люди с диабетом и без
```

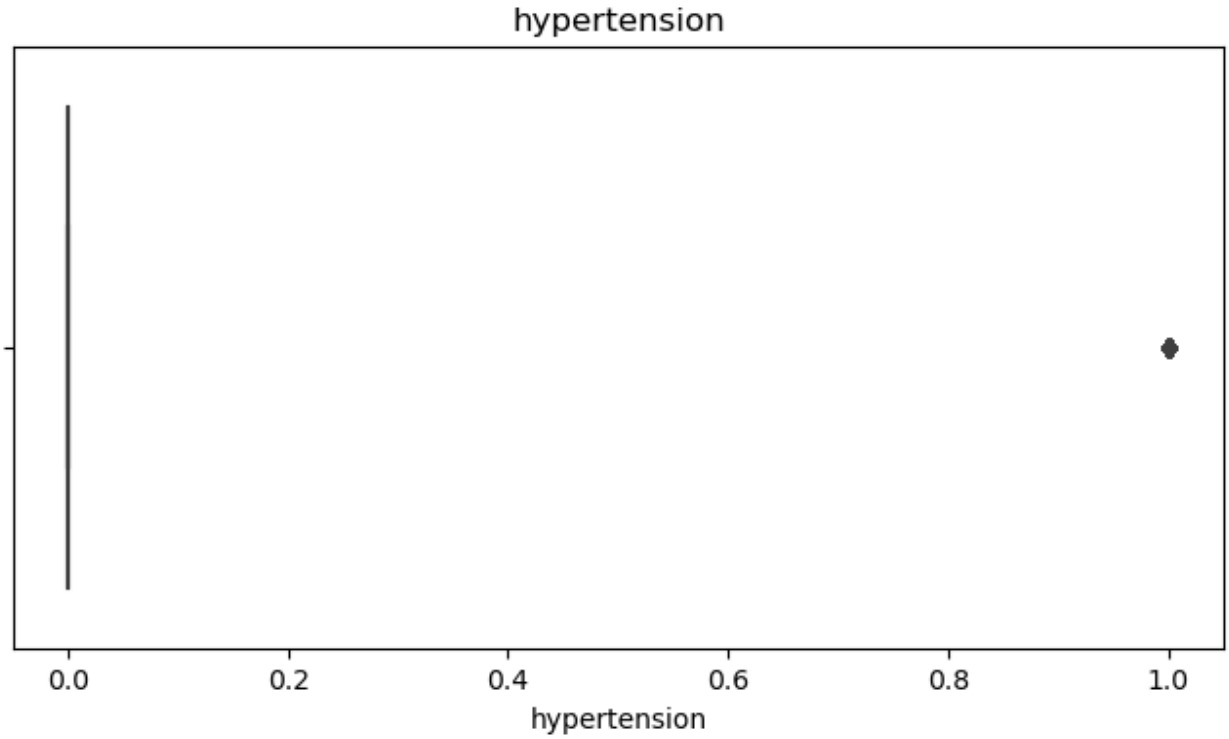
```python
# Гистограммы для числовых переменных
for column in numerical_columns:
    plt.figure(figsize=(8, 4))
    sns.histplot(data=df, x=column, hue="diabetes", element="step", stat=
    plt.title(f'Распределение {column} по статусу диабета')
    plt.show()

# Диаграммы размаха для числовых переменных
for column in numerical_columns:
    plt.figure(figsize=(8, 4))
    sns.boxplot(x='diabetes', y=column, data=df)
    plt.title(f'Ящики с усами для {column} по статусу диабета')
    plt.show()

# Матрица корреляции признаков
plt.figure(figsize=(10, 8))
corr_matrix = df.corr(numeric_only=True)
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Матрица корреляции признаков')
plt.show()
```

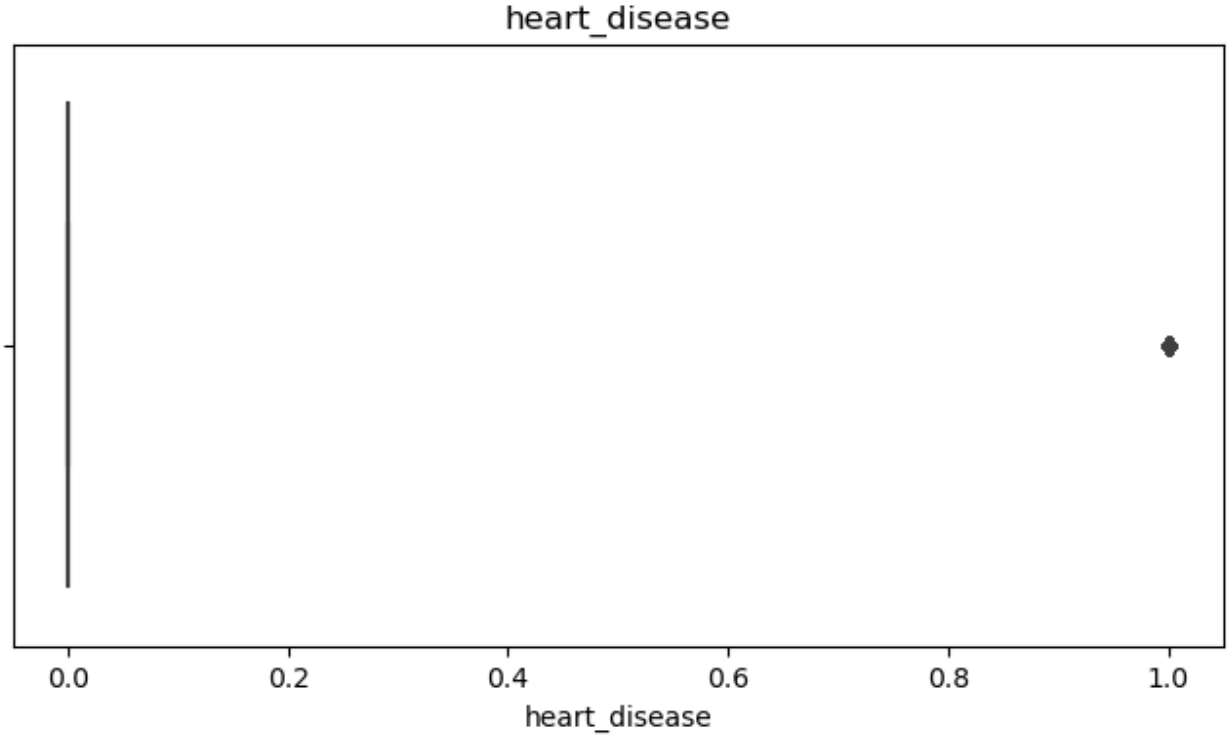
```
Описательный анализ данных:
                 age  hypertension  heart_disease            bmi  \
count  100000.000000  100000.00000  100000.000000  100000.000000
mean       41.885856       0.07485       0.039420      27.320767
std        22.516840       0.26315       0.194593       6.636783
min         0.080000       0.00000       0.000000      10.010000
25%        24.000000       0.00000       0.000000      23.630000
50%        43.000000       0.00000       0.000000      27.320000
75%        60.000000       0.00000       0.000000      29.580000
max        80.000000       1.00000       1.000000      95.690000

         HbA1c_level  blood_glucose_level       diabetes
count  100000.000000        100000.000000  100000.000000
mean        5.527507           138.058060       0.085000
std         1.070672            40.708136       0.278883
min         3.500000            80.000000       0.000000
25%         4.800000           100.000000       0.000000
50%         5.800000           140.000000       0.000000
75%         6.200000           159.000000       0.000000
max         9.000000           300.000000       1.000000
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   gender               100000 non-null  object
 1   age                  100000 non-null  float64
 2   hypertension         100000 non-null  int64
 3   heart_disease        100000 non-null  int64
 4   smoking_history      100000 non-null  object
 5   bmi                  100000 non-null  float64
 6   HbA1c_level          100000 non-null  float64
 7   blood_glucose_level  100000 non-null  int64
 8   diabetes             100000 non-null  int64
dtypes: float64(3), int64(4), object(2)
memory usage: 6.9+ MB
```

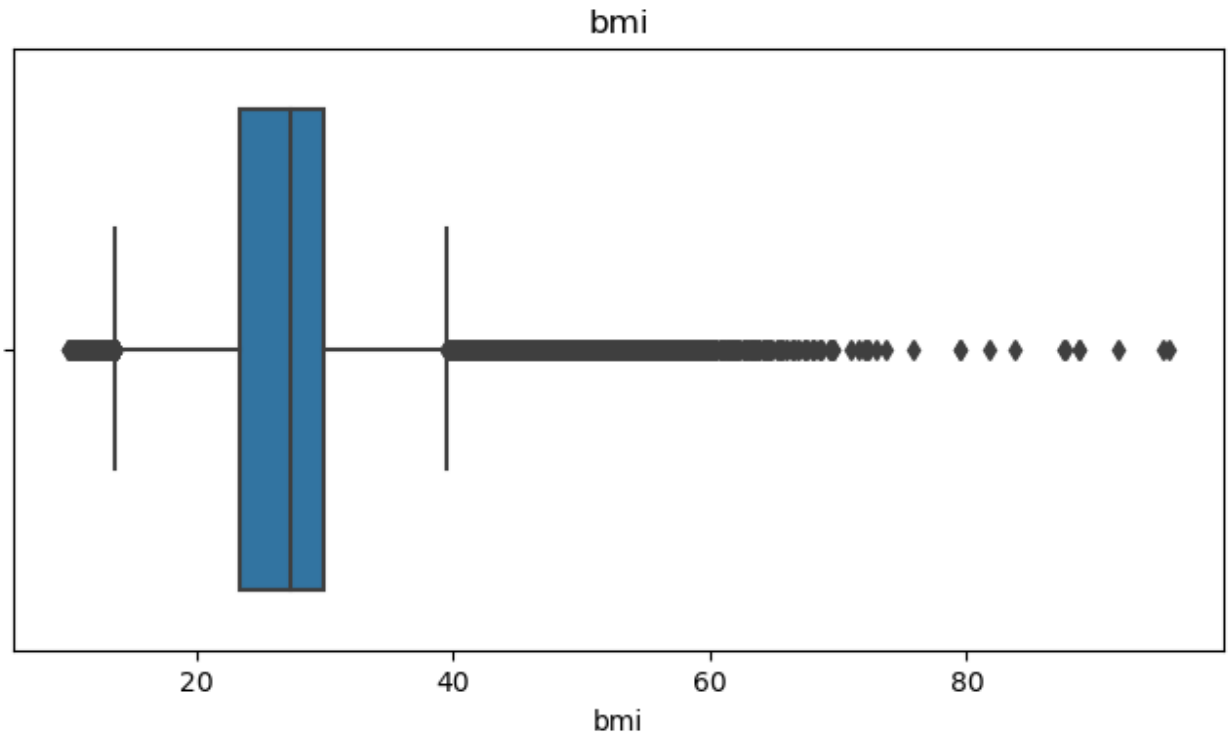
```
None

Предобработка данных: Удаление дубликатов
Удалено дубликатов: 3854

Обработка отсутствующих значений:
Количество NaN до:
gender                 0
age                    0
hypertension           0
heart_disease          0
smoking_history        0
bmi                    0
HbA1c_level            0
blood_glucose_level    0
diabetes               0
dtype: int64
Количество NaN после:
gender                 0
age                    0
hypertension           0
heart_disease          0
smoking_history        0
bmi                    0
HbA1c_level            0
blood_glucose_level    0
diabetes               0
dtype: int64
```

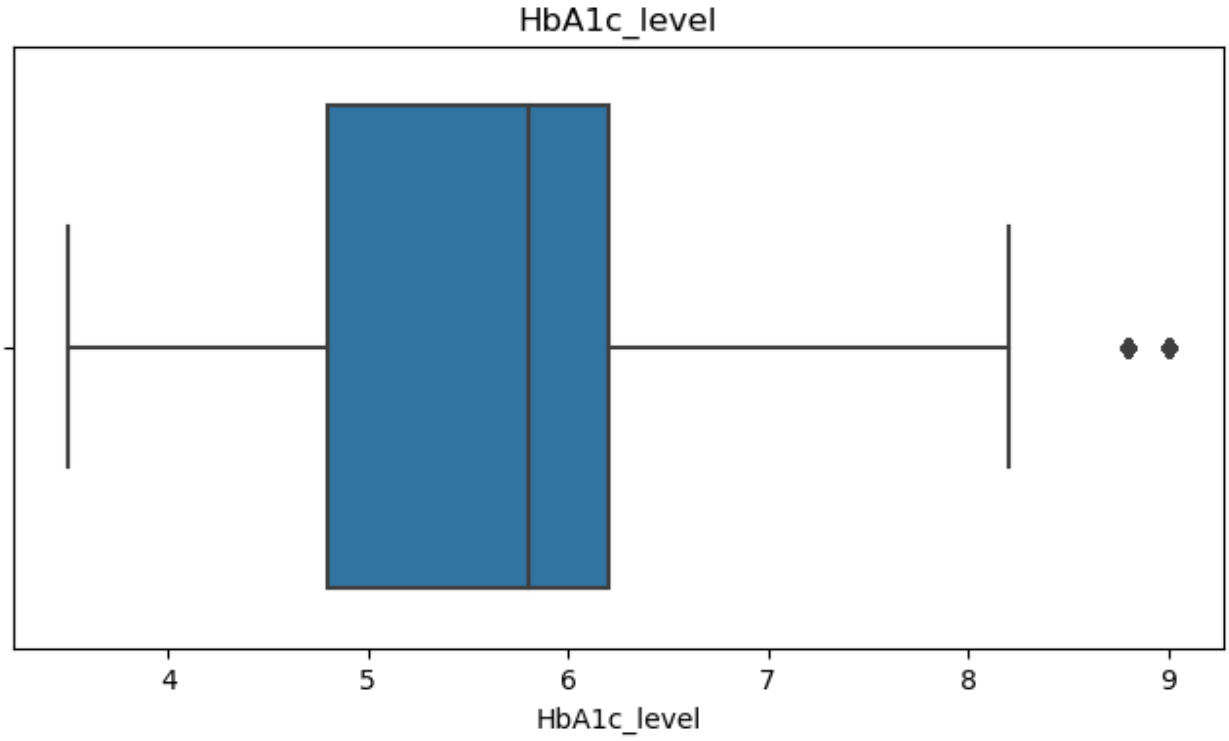

age
hypertension
heart_disease
bmi
HbA1c_level
blood_glucose_level
diabetes

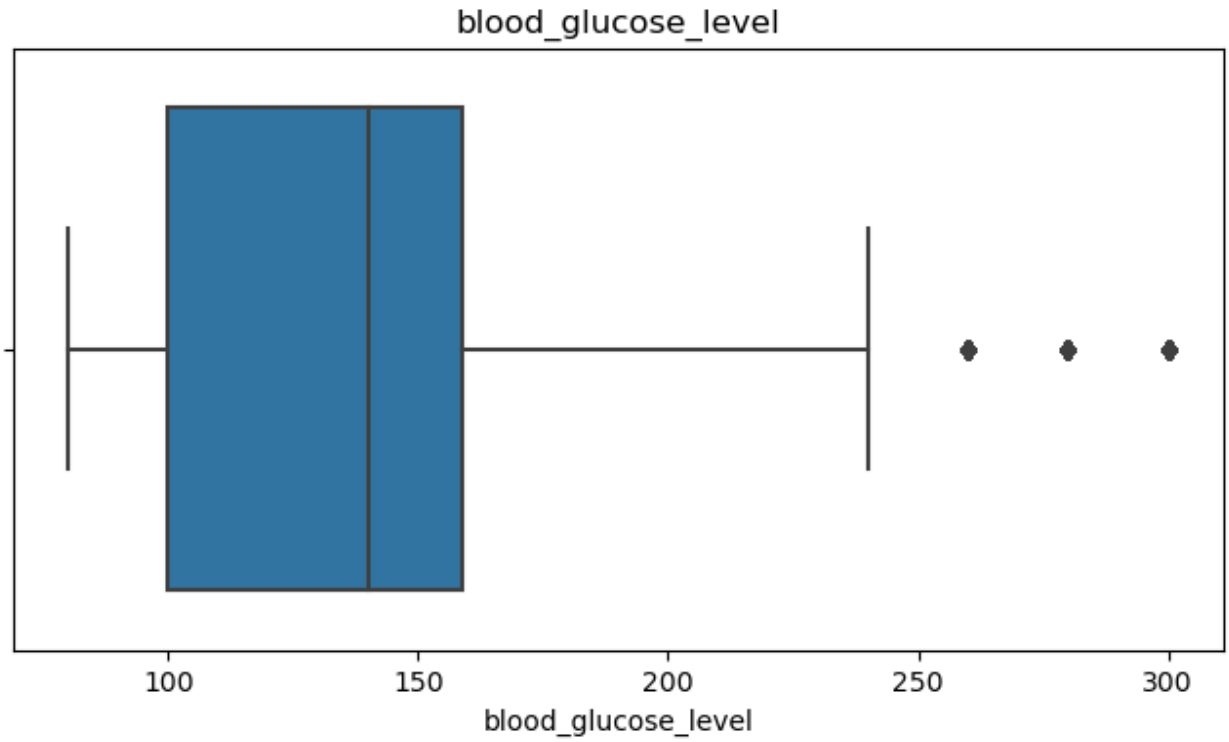

age
0
10
20
30
40
50
60
70
80
age

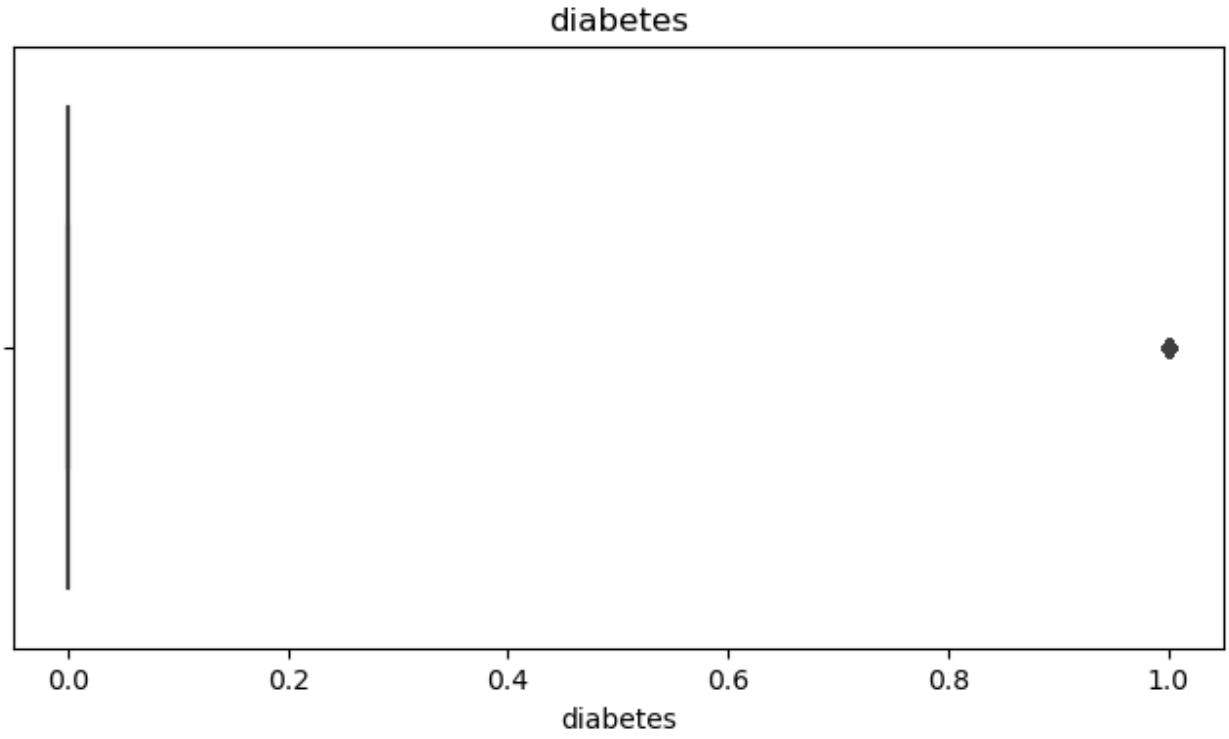

hypertension
0.0
0.2
0.4
0.6
0.8
1.0
hypertension

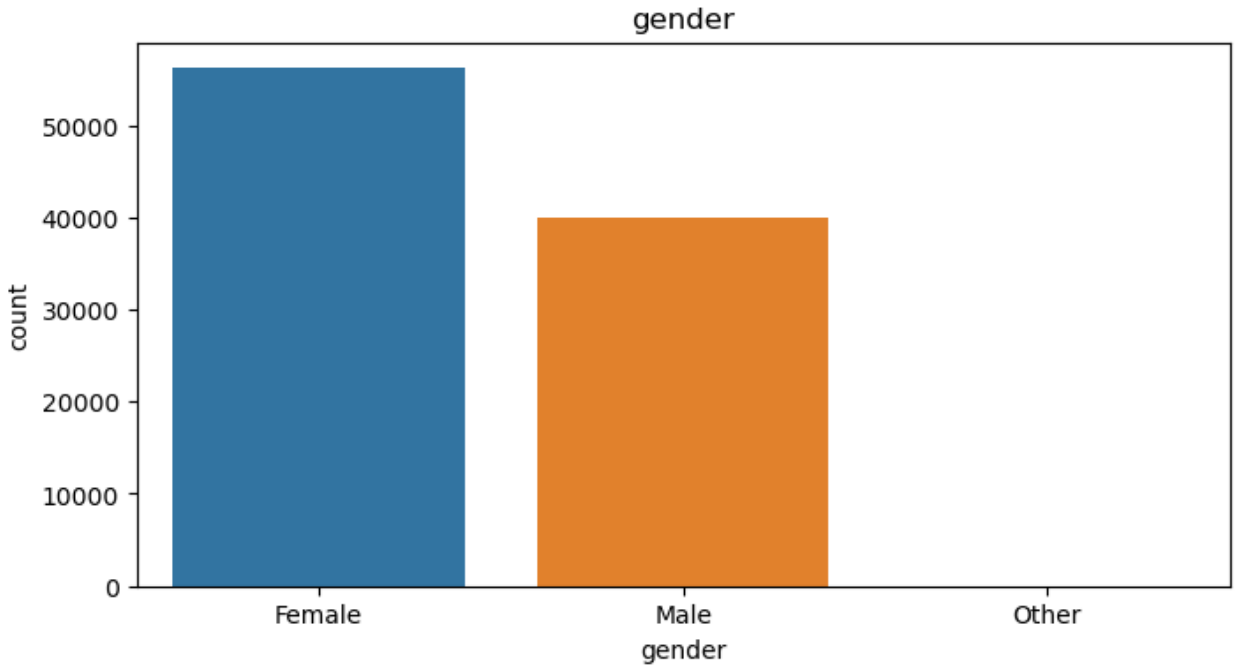

heart_disease
0.0
0.2
0.4
0.6
0.8
1.0
heart_disease

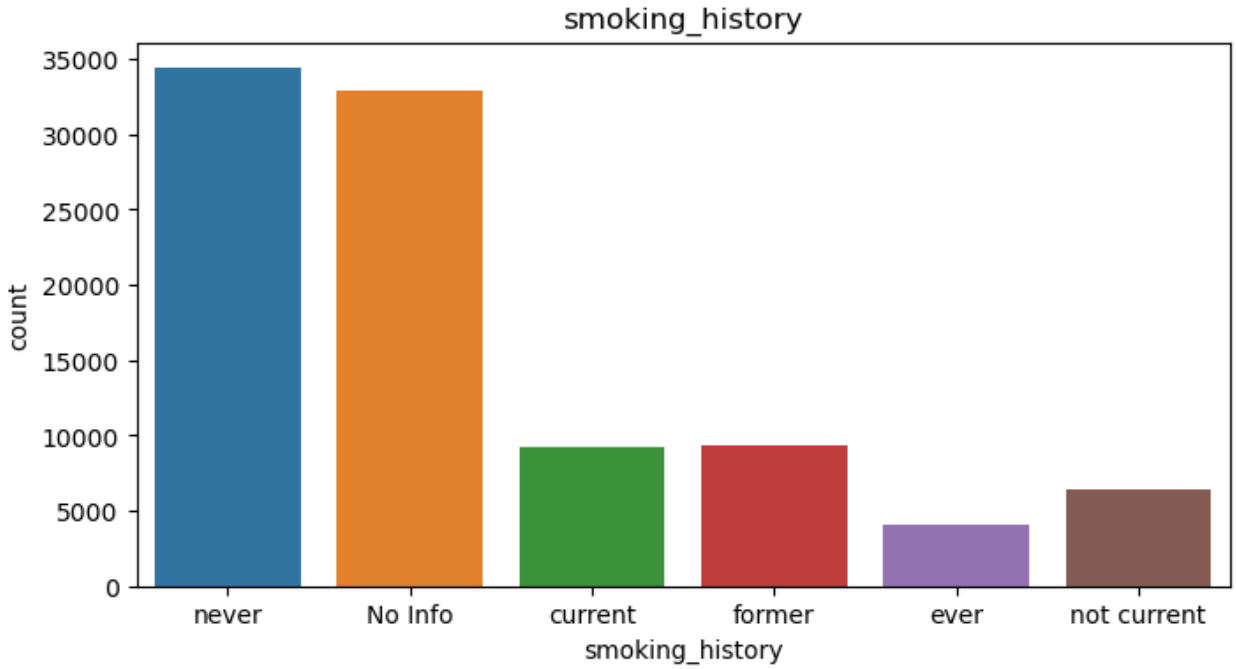

bmi
20
40
60
80
bmi

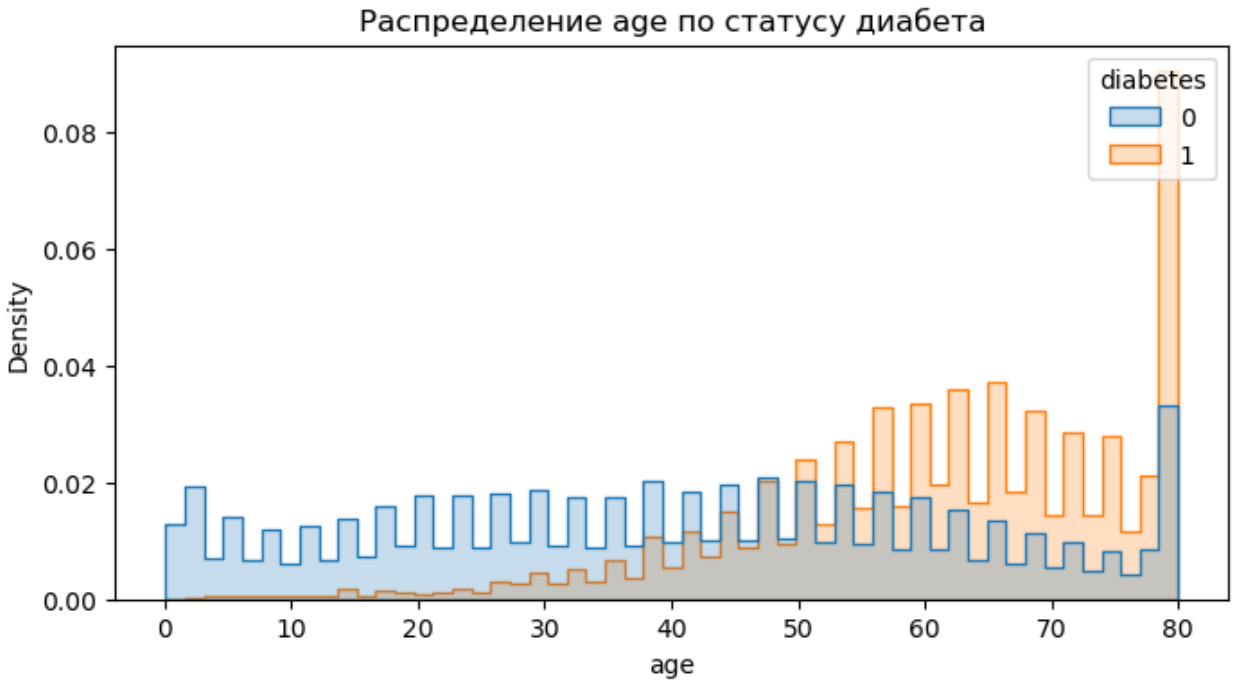

HbA1c_level
4
5
6
7
8
9
HbA1c_level


blood_glucose_level
100
150
200
250
300
blood_glucose_level

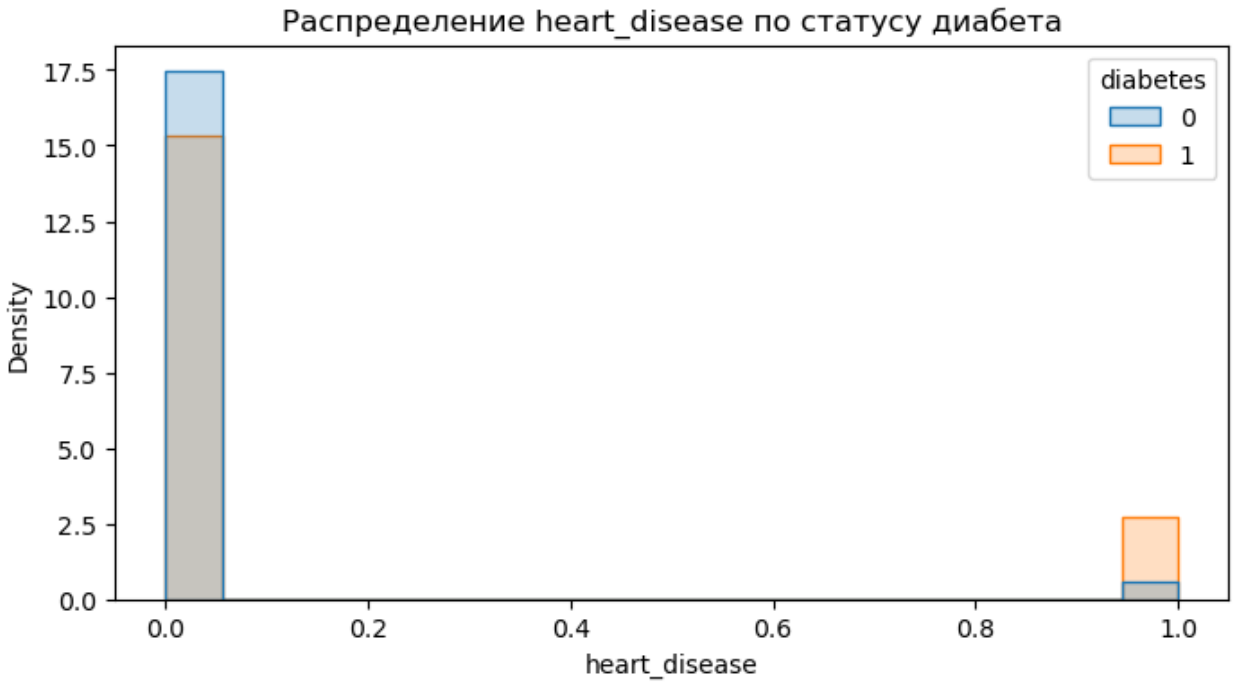

diabetes
0.0
0.2
0.4
0.6
0.8
1.0
diabetes

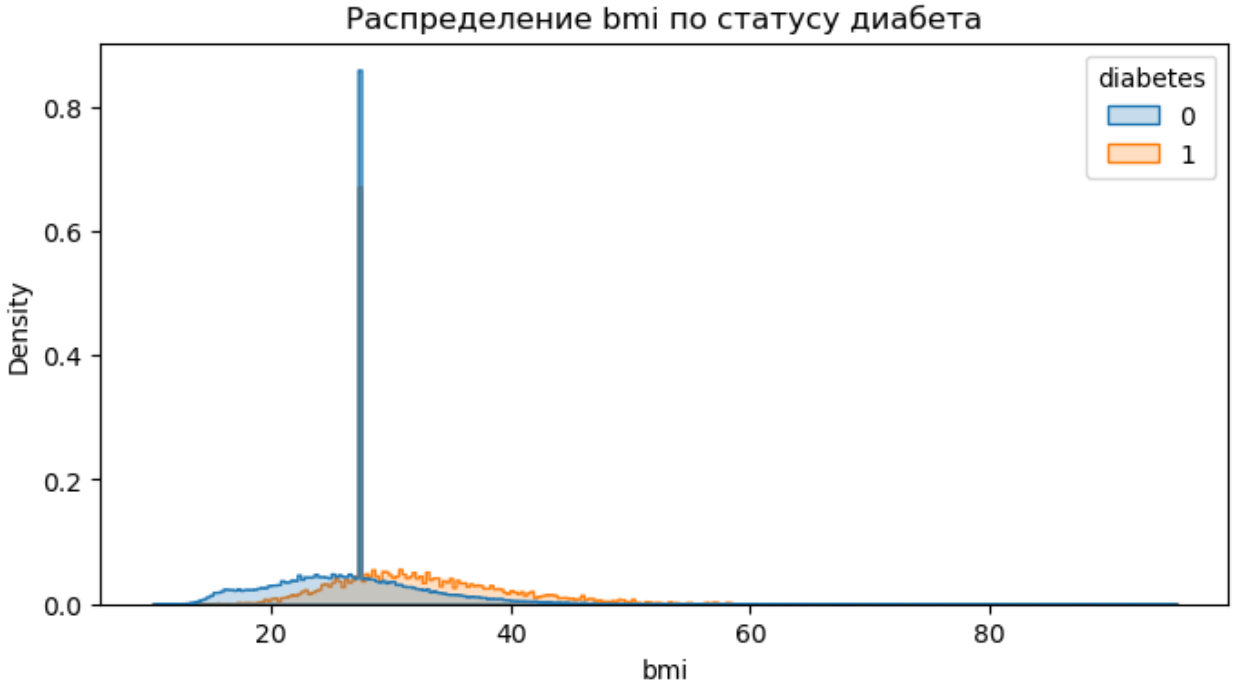

gender
count
50000
40000
30000
20000
10000
0
Female
Male
Other
gender

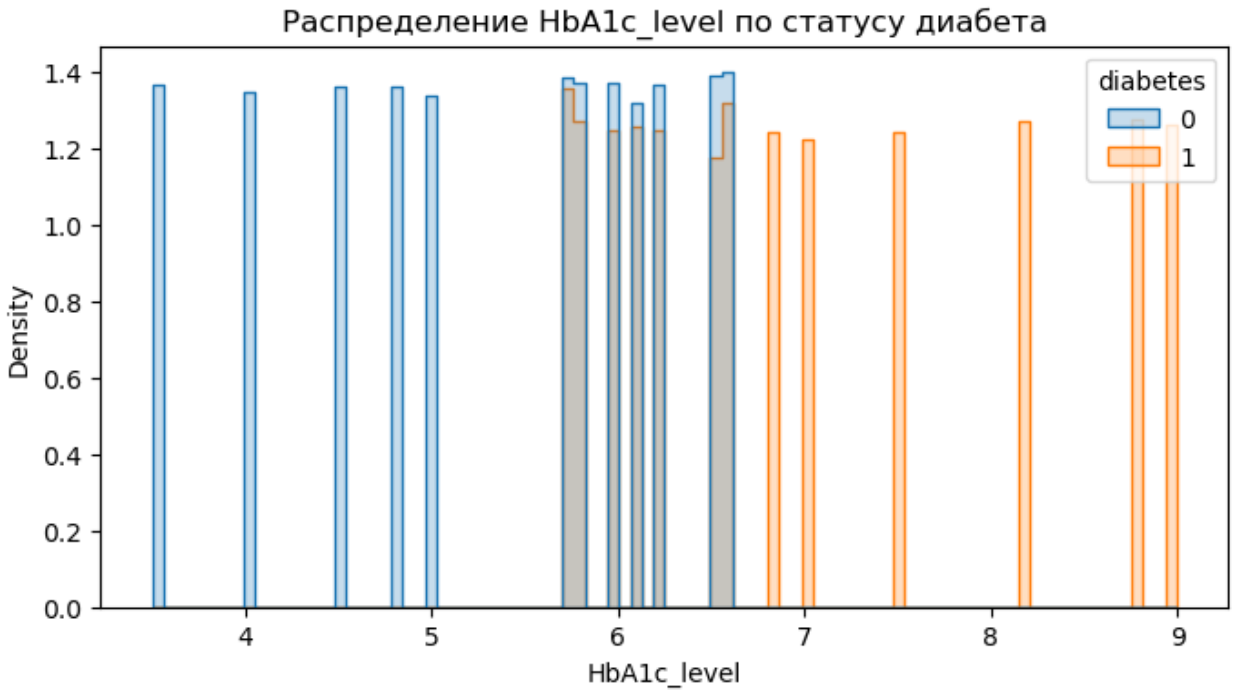

smoking_history
count
35000
30000
25000
20000
15000
10000
5000
0
never
No Info
current
former
ever
not current
smoking_history

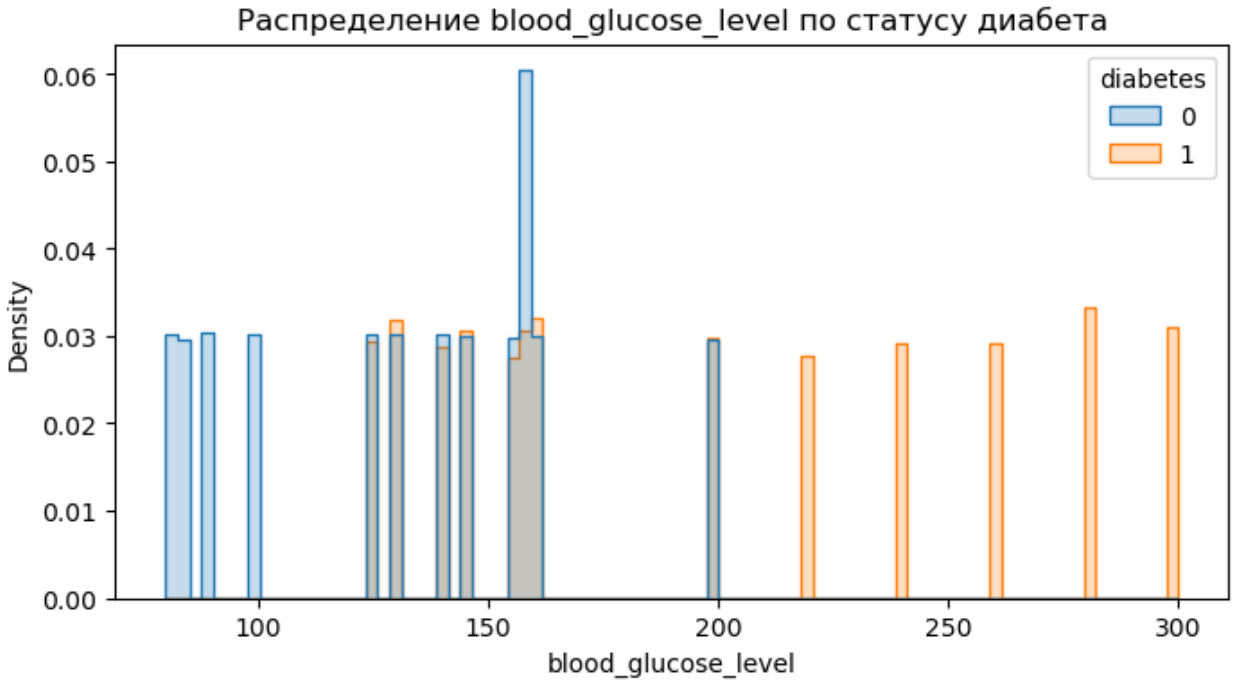

Распределение age по статусу диабета
diabetes
0
1
Density
0.00
0.02
0.04
0.06
0.08
0
10
20
30
40
50
60
70
80
age

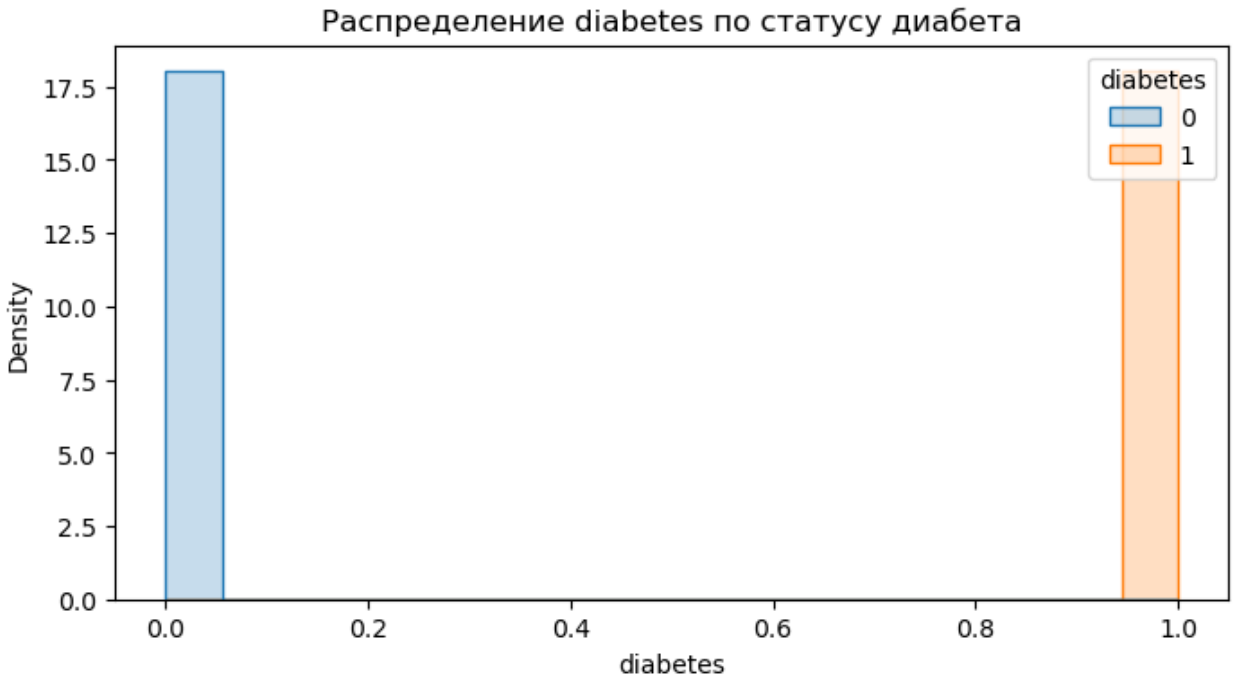

Распределение hypertension по статусу диабета
diabetes
0
1
Density
0
2
4
6
8
10
12
14
16
0.0
0.2
0.4
0.6
0.8
1.0
hypertension

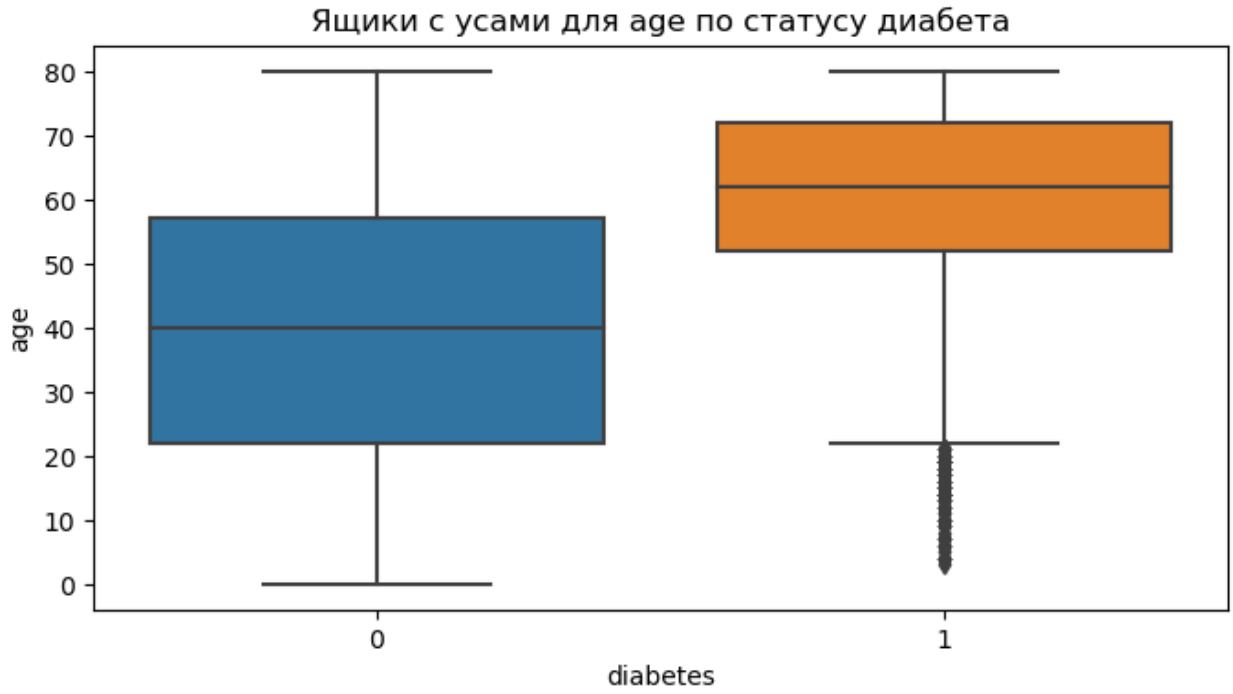

Распределение heart_disease по статусу диабета
diabetes
0
1
Density
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
heart_disease

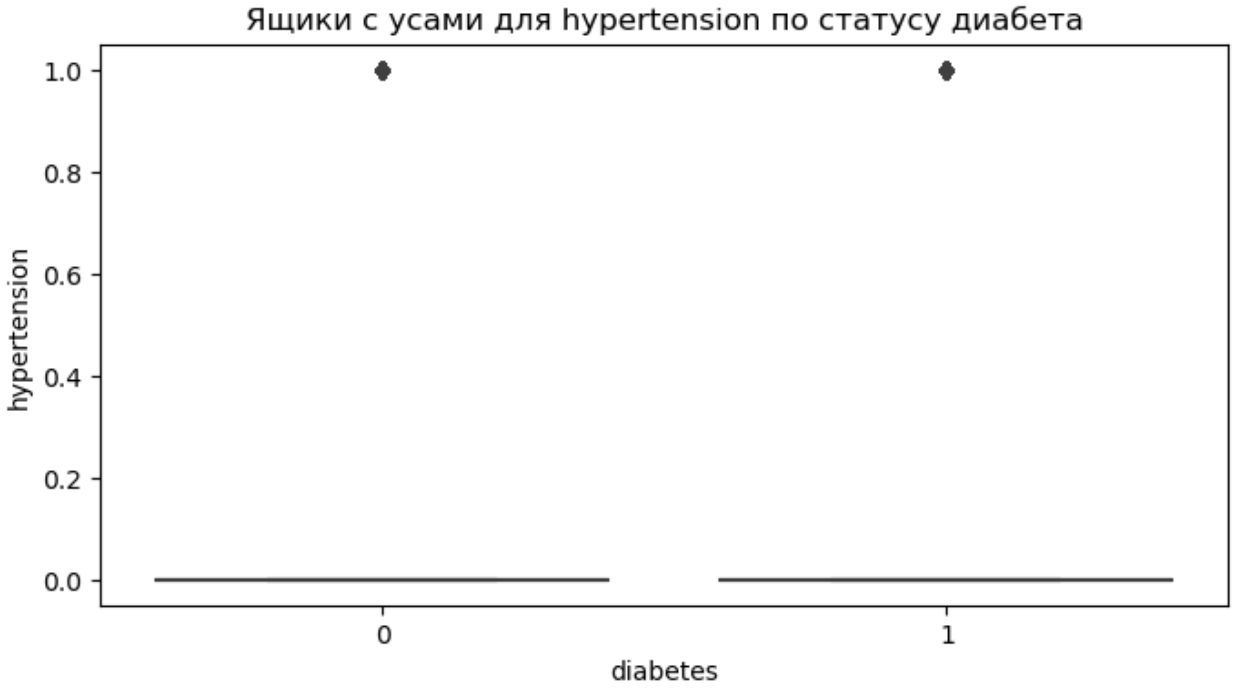

Распределение bmi по статусу диабета
diabetes
0
1
Density
0.8
0.6
0.4
0.2
0.0
20
40
60
80
bmi

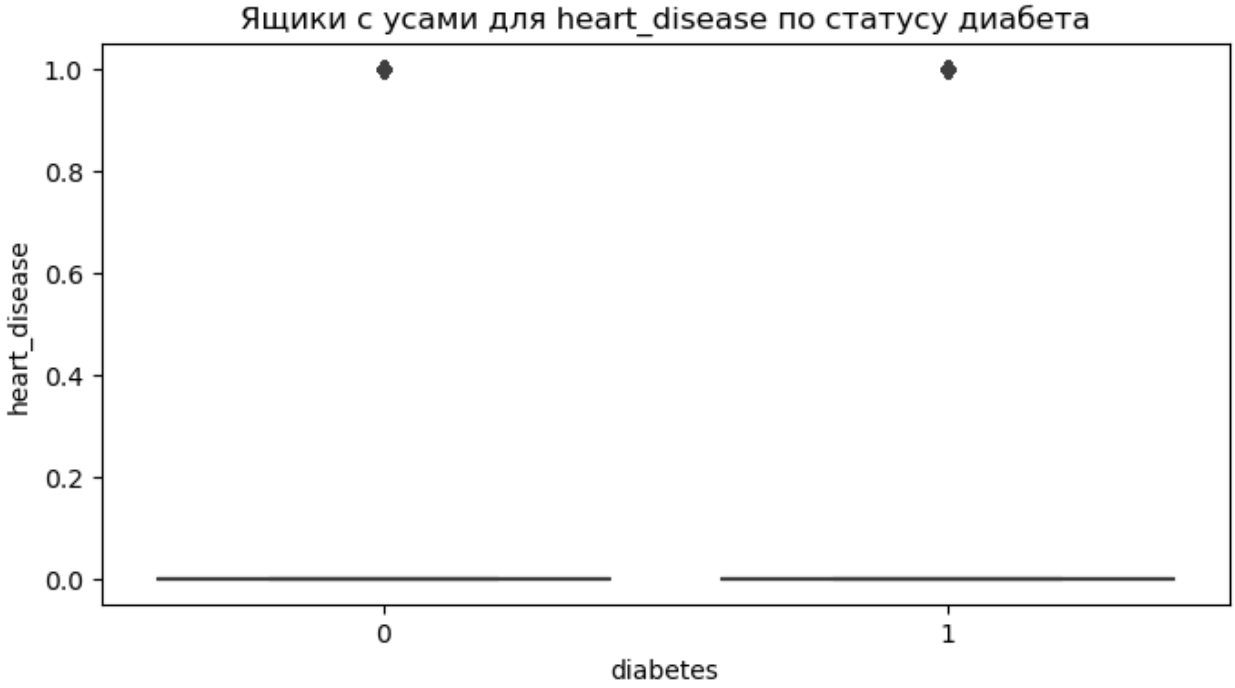

Распределение HbA1c_level по статусу диабета
diabetes
0
1
Density
HbA1c_level
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
4
5
6
7
8
9

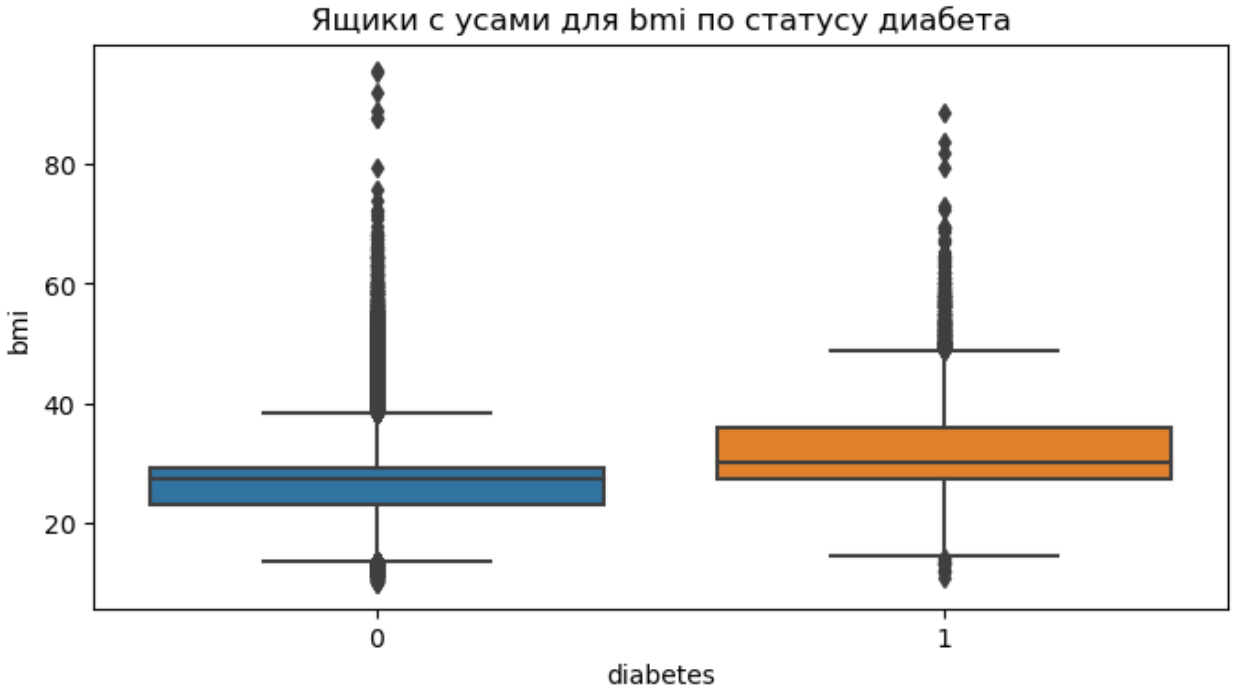

Распределение blood_glucose_level по статусу диабета
diabetes
0
1
Density
blood_glucose_level
0.00
0.01
0.02
0.03
0.04
0.05
0.06
100
150
200
250
300

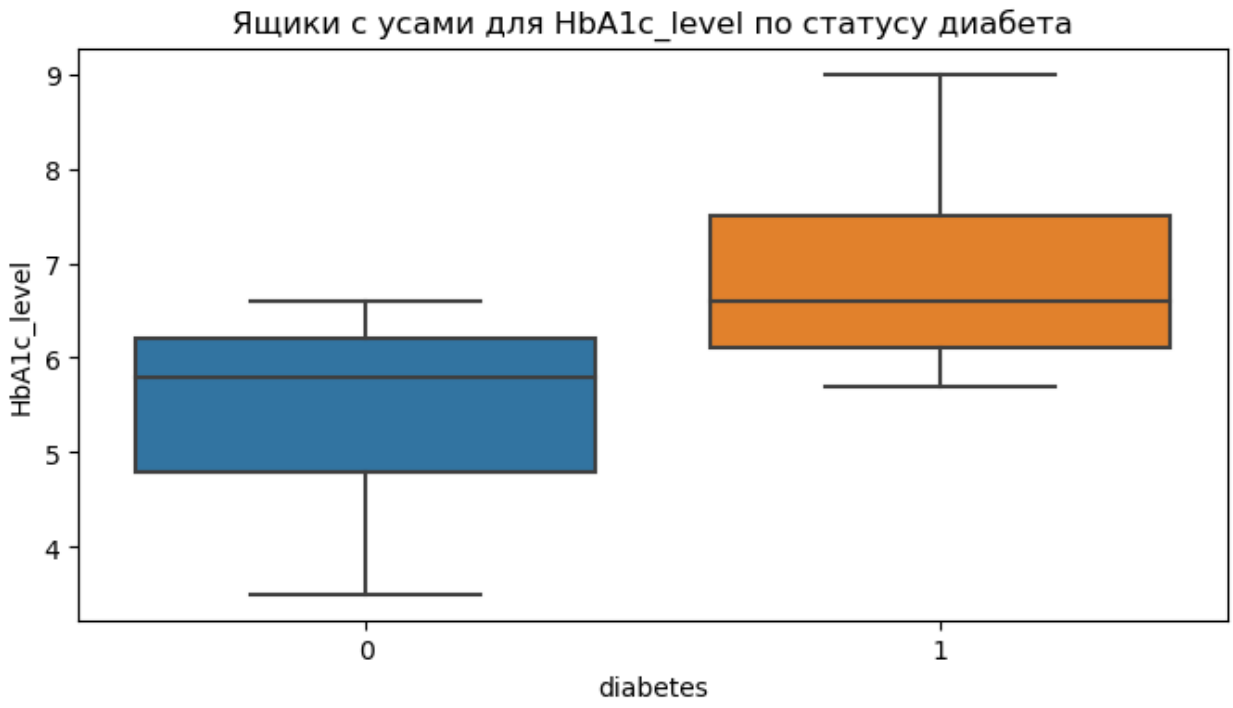

Распределение diabetes по статусу диабета
diabetes
0
1
Density
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
diabetes

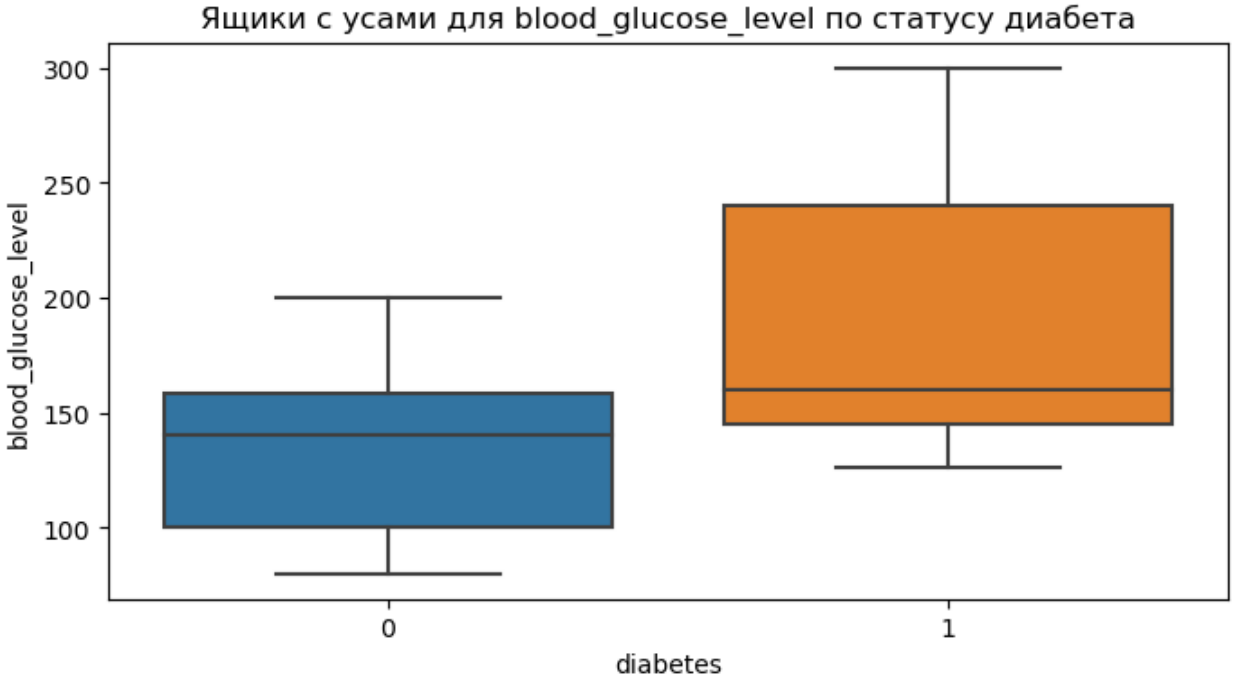

Ящики с усами для age по статусу диабета
age
80
70
60
50
40
30
20
10
0
0
1
diabetes

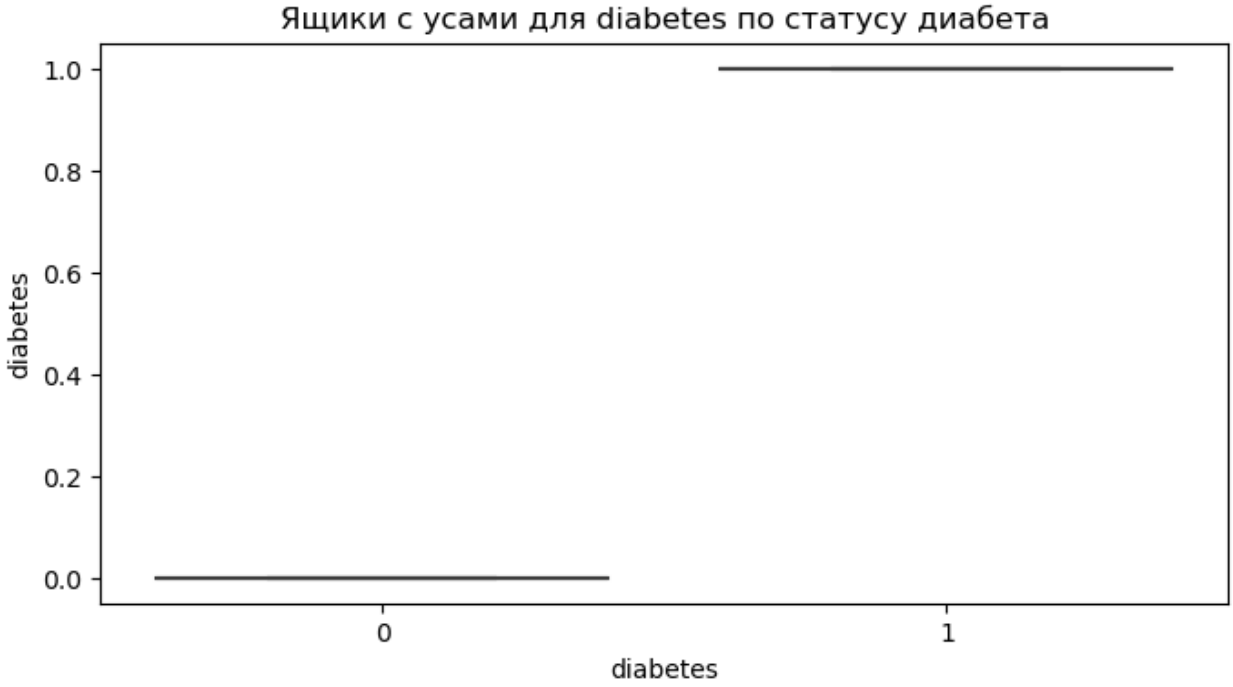

Ящики с усами для hypertension по статусу диабета
hypertension
1.0
0.8
0.6
0.4
0.2
0.0
0
1
diabetes

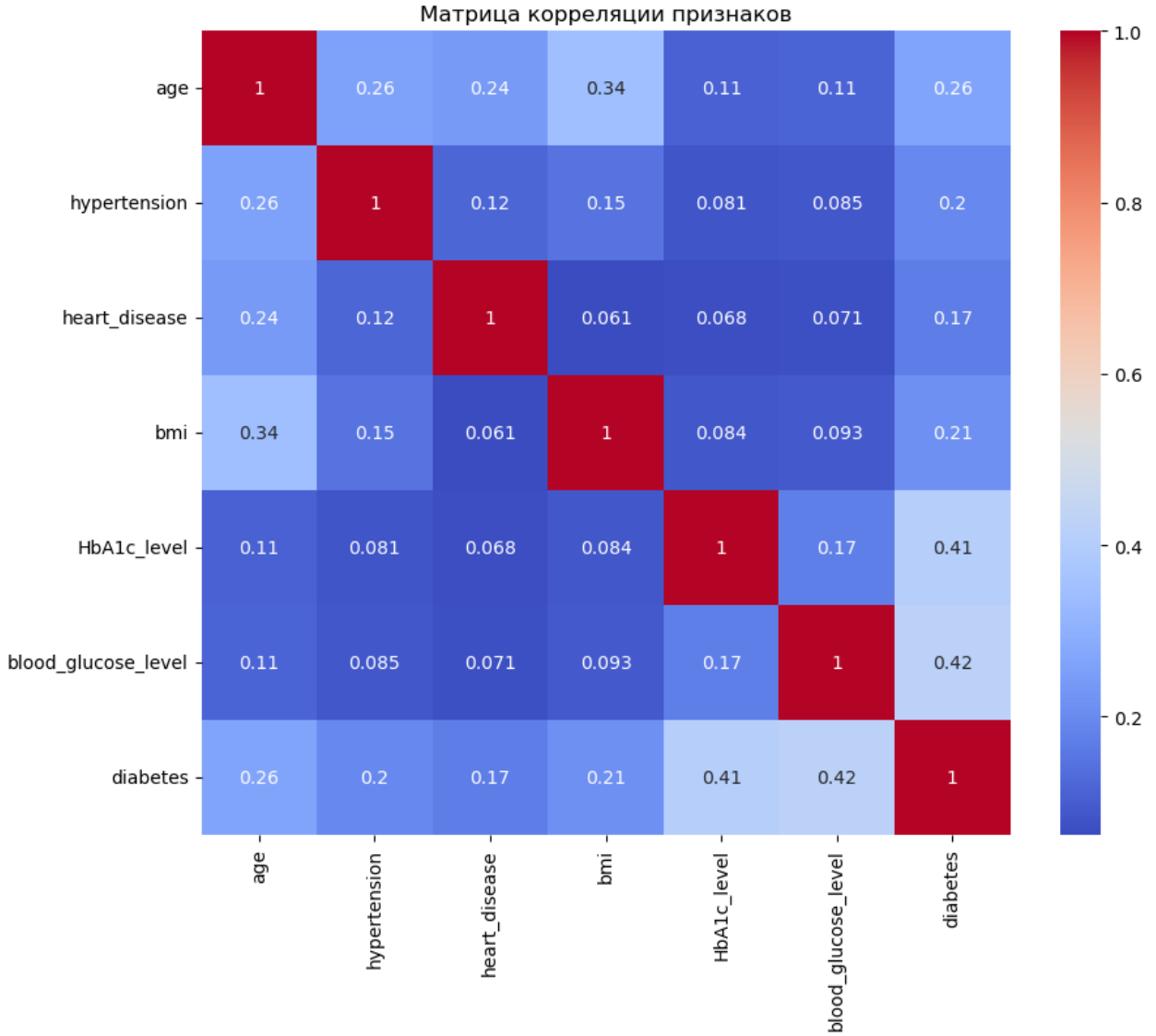

Ящики с усами для heart_disease по статусу диабета
heart_disease
1.0
0.8
0.6
0.4
0.2
0.0
0
1
diabetes


Ящики с усами для bmi по статусу диабета
bmi
80
60
40
20
0
1
diabetes


Ящики с усами для HbA1c_level по статусу диабета
HbA1c_level
9
8
7
6
5
4
0
1
diabetes


Ящики с усами для blood_glucose_level по статусу диабета
blood_glucose_level
300
250
200
150
100
0
1
diabetes


Ящики с усами для diabetes по статусу диабета
diabetes
1.0
0.8
0.6
0.4
0.2
0.0
0
1
diabetes


Матрица корреляции признаков
age
hypertension
heart_disease
bmi
HbA1c_level
blood_glucose_level
diabetes
1 0.26 0.24 0.34 0.11 0.11 0.26
0.26 1 0.12 0.15 0.081 0.085 0.2
0.24 0.12 1 0.061 0.068 0.071 0.17
0.34 0.15 0.061 1 0.084 0.093 0.21
0.11 0.081 0.068 0.084 1 0.17 0.41
0.11 0.085 0.071 0.093 0.17 1 0.42
0.26 0.2 0.17 0.21 0.41 0.42 1
1.0
0.8
0.6
0.4
0.2

In [ ]: