|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

**Факультет «Информатика и системы управления»**

**Кафедра «Системы обработки информации и управления»**

Отчет по рубежному контролю №1

**«Технологии разведочного анализа и обработки данных»**

по дисциплине «Технологии машинного обучения»

Вариант №28

Выполнил:

студент группы ИУ5Ц-84Б

Перевощиков Н.Д.

подпись, дата

Проверил:

к.т.н., доц., Гапанюк Ю.Е.

подпись, дата

2024 г.

**СОДЕРЖАНИЕ ОТЧЕТА**

[1. Примечание 3](#_Toc162194261)

[2. Дополнительные требования по группам 3](#_Toc162194262)

[3. Листинг 3](#_Toc162194263)

[3.1. Подключение библиотеки и импорт данных 3](#_Toc162194264)

[3.2. Изучение данных 4](#_Toc162194265)

[3.3. Анализ выбросов (Ящик с усами) 6](#_Toc162194266)

[3.4. Пропуски 7](#_Toc162194267)

[3.5. Дубликаты 7](#_Toc162194268)

[3.6. Обработка категориальных признаков 7](#_Toc162194269)

[3.7. Создание нового признака 8](#_Toc162194270)

[3.7.1. Группировка по возрастам 8](#_Toc162194271)

[3.7.2. Группировка по уровню частоты сердечных заболеваний 9](#_Toc162194272)

[3.8. Частота сердечных сокращений в разрезе пола 10](#_Toc162194273)

[3.9. Распределение пола и наличия сердечных заболеваний 11](#_Toc162194274)

## Примечание

Если в Вашем наборе данных отсутствуют данные, необходимые для решения задачи, создайте их искусственно. Например, если отсутствуют категориальные признаки, создайте категориальный признак на основе числового. Если отсутствуют пропуски, замените на пропуски часть значений в одном или нескольких признаках.

Также Вы можете дополнительно использовать датасеты, содержащие необходимые данные, например использовать дополнительный датасет, содержащий пропуски.

## Дополнительные требования по группам

Для студентов группы ИУ5-64Б, ИУ5Ц-84Б - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".

**Задача №4**

Для заданного набора данных постройте основные графики, входящие в этап разведочного анализа данных. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Какие графики Вы построили и почему? Какие выводы о наборе данных Вы можете сделать на основании построенных графиков?

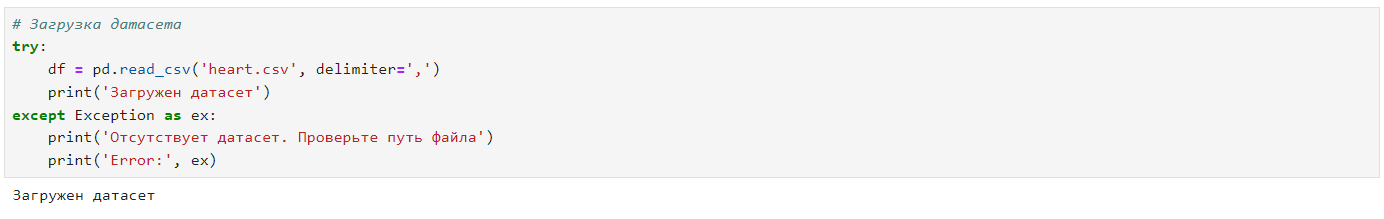
**Наборы данных:**

https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

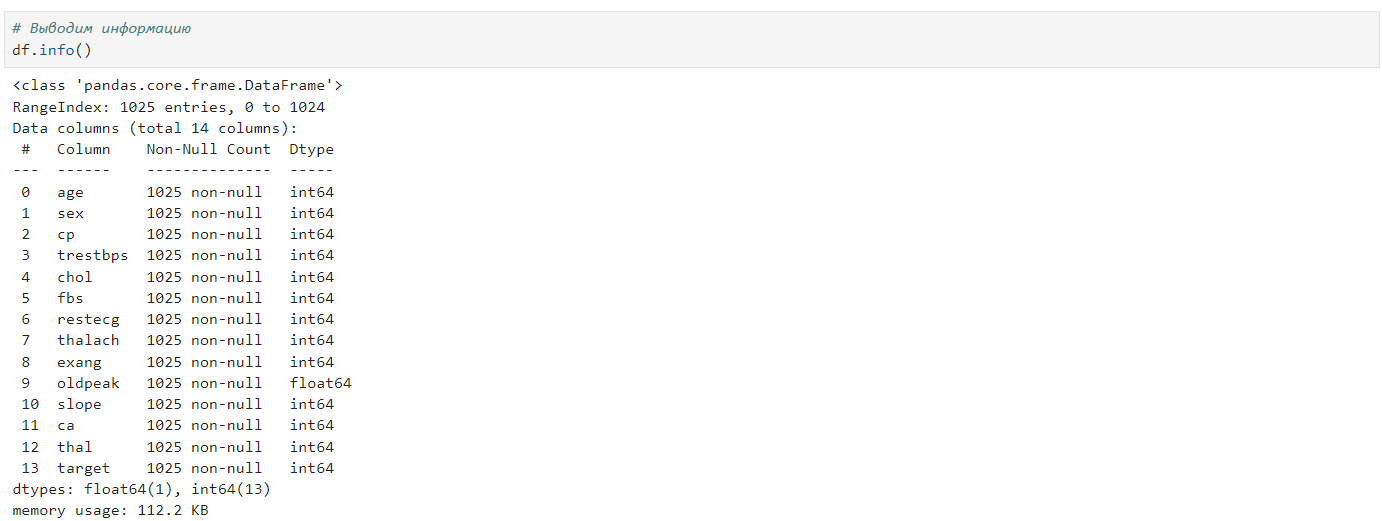
## Листинг

## Подключение библиотеки и импорт данных

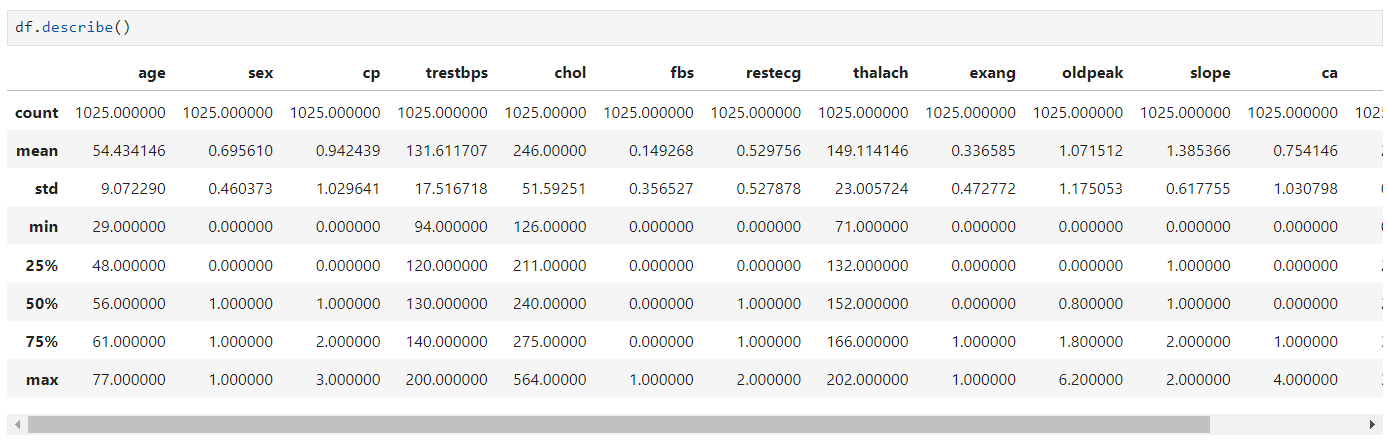
****

****

## Изучение данных

****

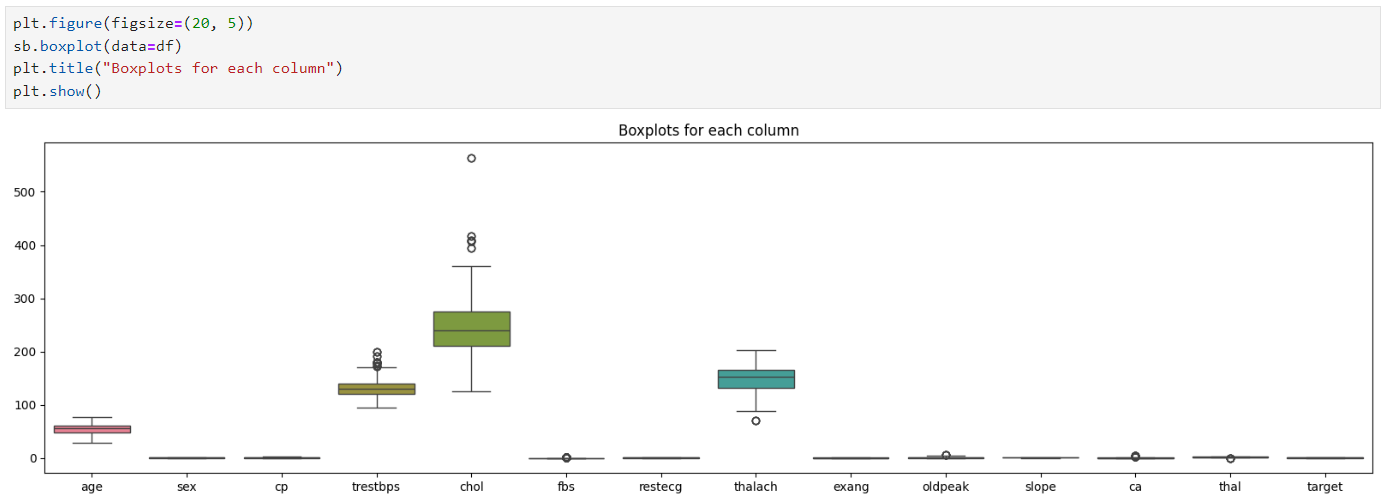
****

****

##### Анализ описательной статистики

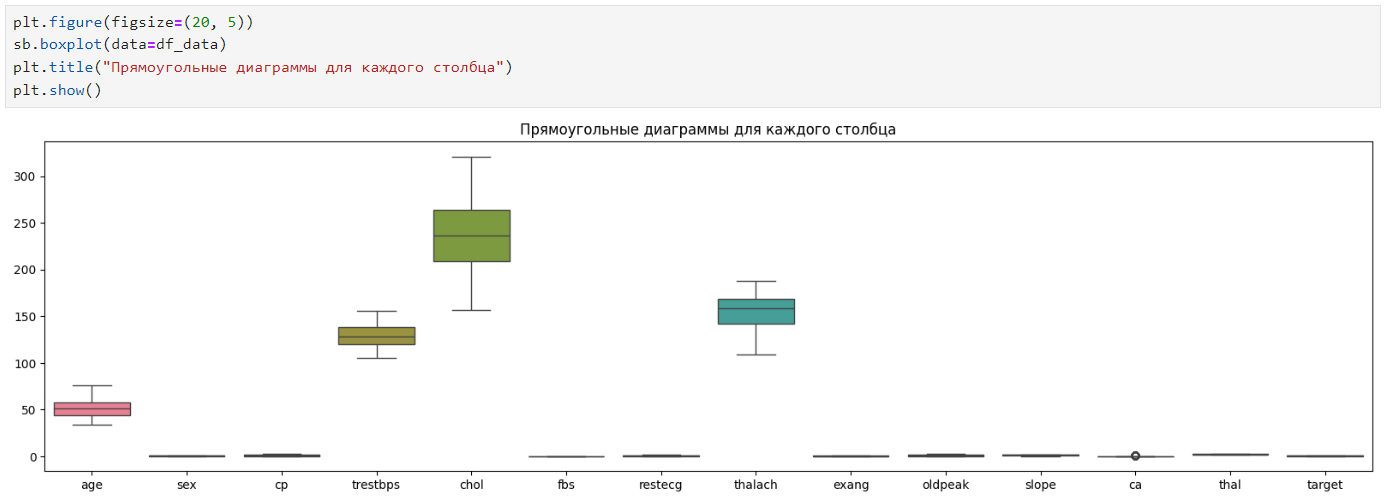
* **age:** Средний возраст составляет около 54 лет. Минимальный возраст в данных - 29 лет, а максимальный - 77 лет.
* **sex:** Примерно 70% данных относятся к мужчинам.
* **cp:** Значения от 0 до 3, представляющие различные типы боли в груди.
* **trestbps:** Среднее кровяное давление около 131.6 мм рт.ст. Минимальное значение - 94 мм рт.ст., а максимальное - 200 мм рт.ст.
* **chol:** Средний уровень холестерина составляет 246 мг/дл. Минимальное значение - 126 мг/дл, а максимальное - 564 мг/дл.
* **fbs:** Примерно 15% людей имеют уровень сахара в крови натощак выше 120 мг/дл.
* **restecg:** Значения от 0 до 2, представляющие результаты электрокардиографии в состоянии покоя.
* **thalach:** Средняя максимальная частота сердечных сокращений около 149 ударов в минуту. Минимальное значение - 71 удар в минуту, а максимальное - 202 удара в минуту.
* **exang:** Значения 1 и 0, указывающие на наличие или отсутствие стенокардии, вызванной физической нагрузкой.
* **oldpeak:** Среднее значение депрессии ST около 0.34. Минимальное значение - 0, а максимальное - 6.2.
* **slope:** Значения от 0 до 2, представляющие наклон пика упражнений ST при нагрузке.
* **ca:** Значения от 0 до 3, указывающие на количество крупных сосудов.
* **thal:** Значения от 1 до 3, представляющие различные состояния сердечной недостаточности.
* **target:** Значения 1 и 0, указывающие на наличие или отсутствие сердечного заболевания.

## Анализ выбросов (Ящик с усами)

****

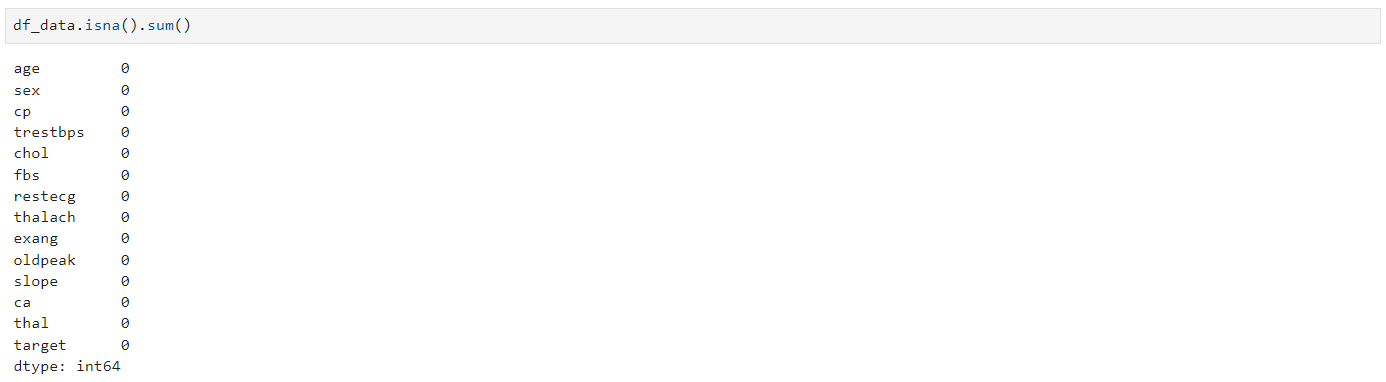
Выбросы имеются у колонок trestbps, chol, fbs, thalach, oldpeak, ca, thal.

****

****

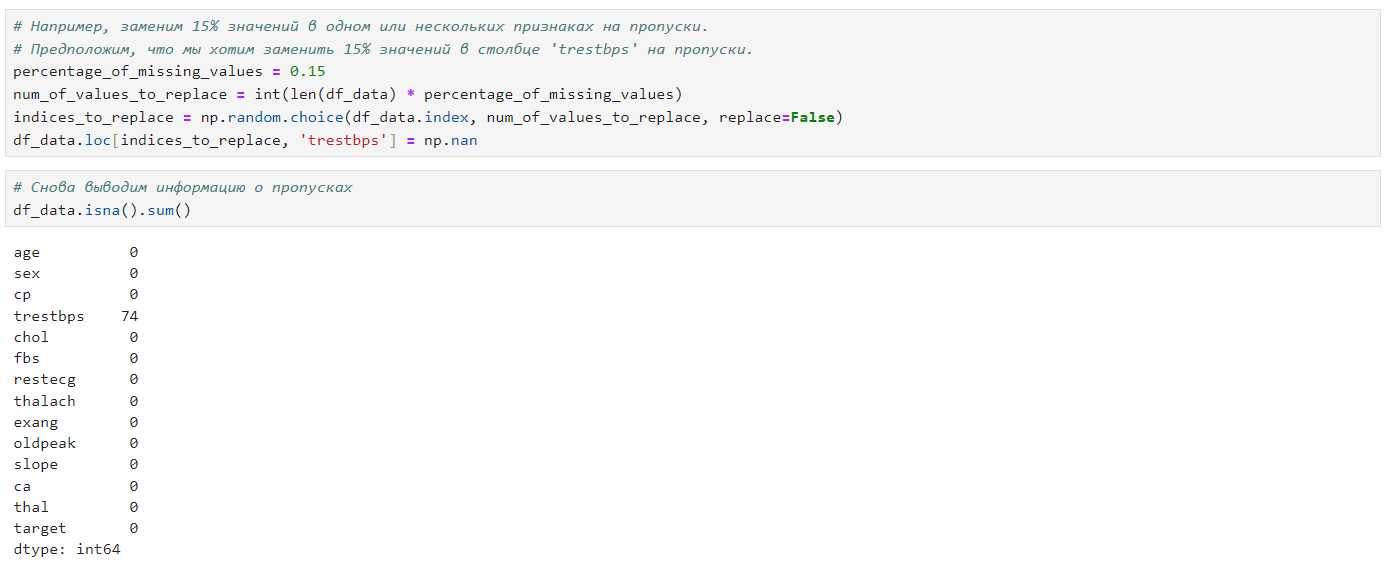
Результат получился нормальной диаграммой с минимальным выбросом.

## Пропуски

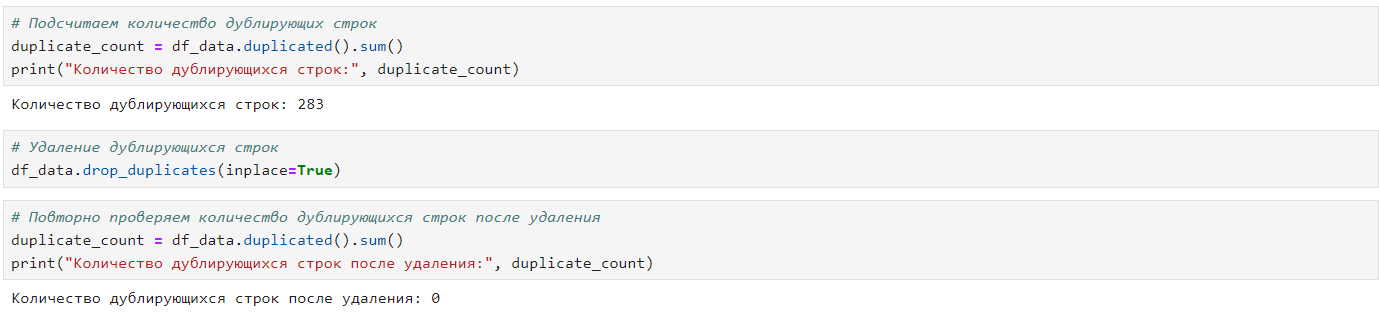
****

Пропусков нет, поэтому создаем искусственно по условию задания.

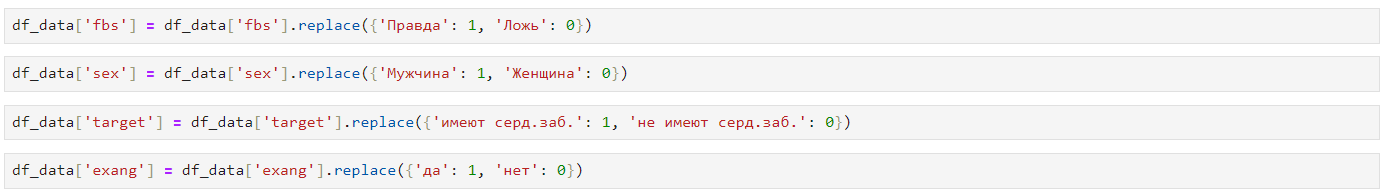
Например, заменим 15% значений в одном или нескольких признаках на пропуски. Предположим, что мы хотим заменить 15% значений в столбце 'trestbps' на пропуски.

****

## Дубликаты

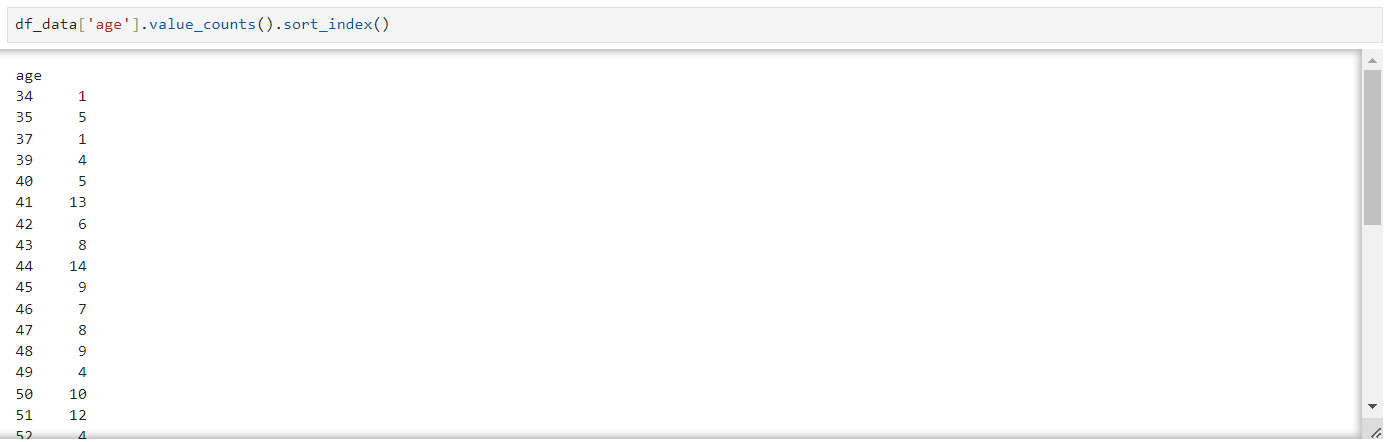
****

## Обработка категориальных признаков

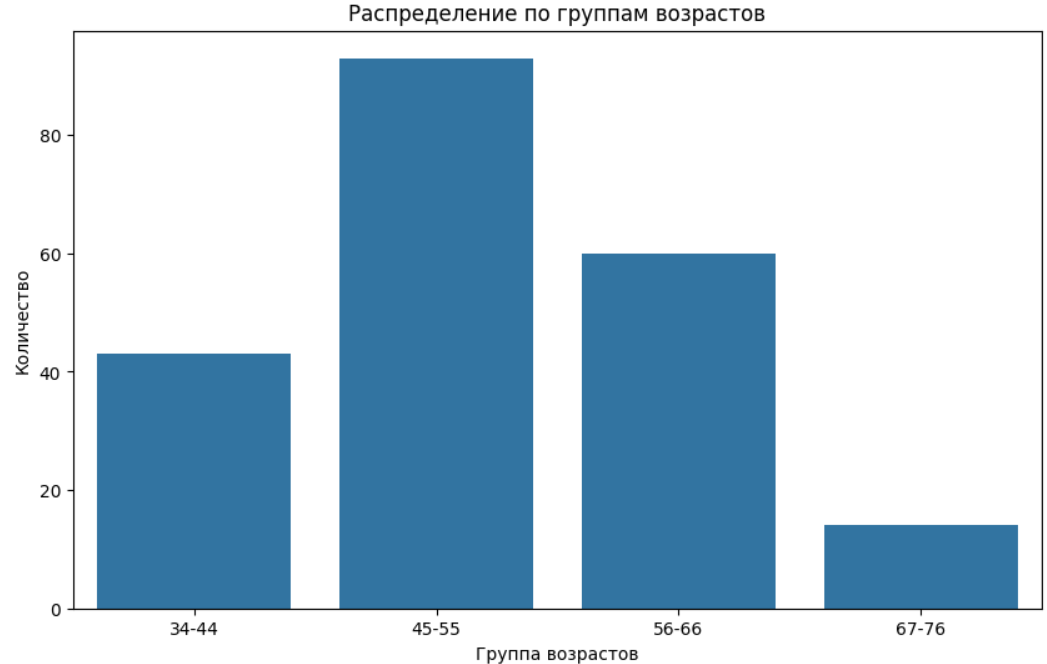
****

## Создание нового признака

## Группировка по возрастам

****

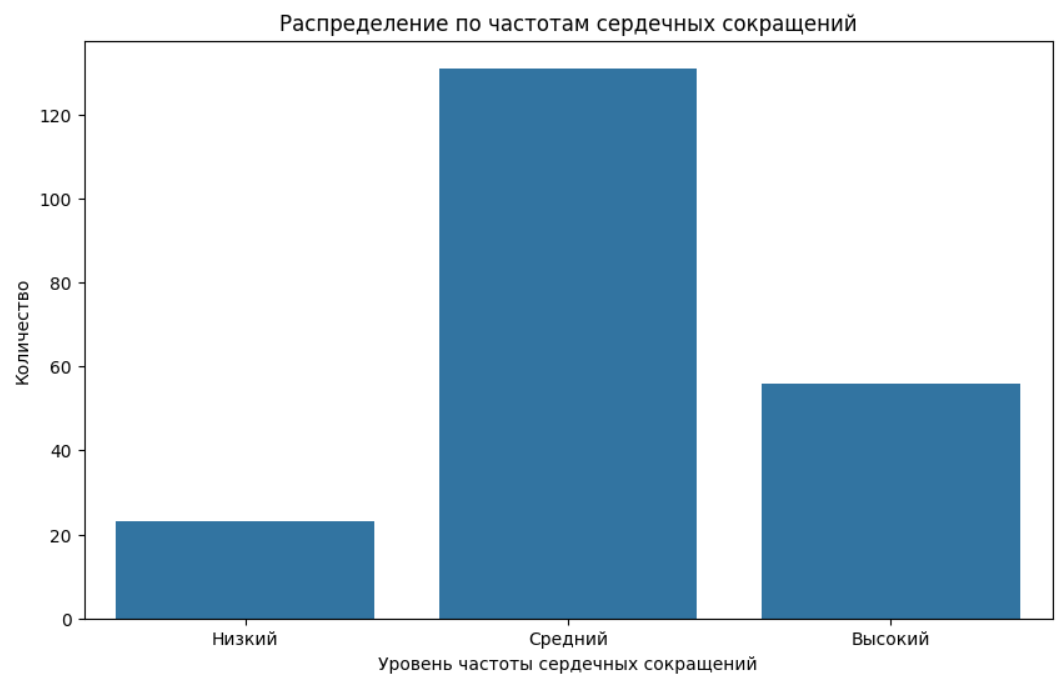
****

****

По графику можно сказать, что наибольшее количество людей имеют возраст в диапазоне от 45 до 55 лет, в то время как наименьшее количество людей приходится на возраст от 67 до 76 лет.

## Группировка по уровню частоты сердечных заболеваний

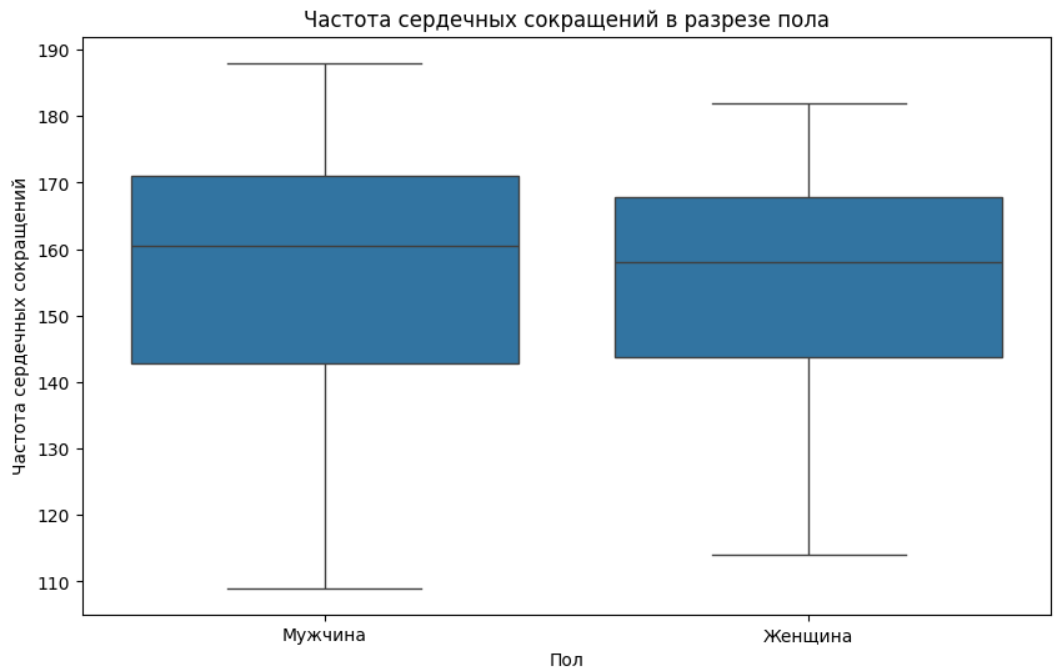
****

****

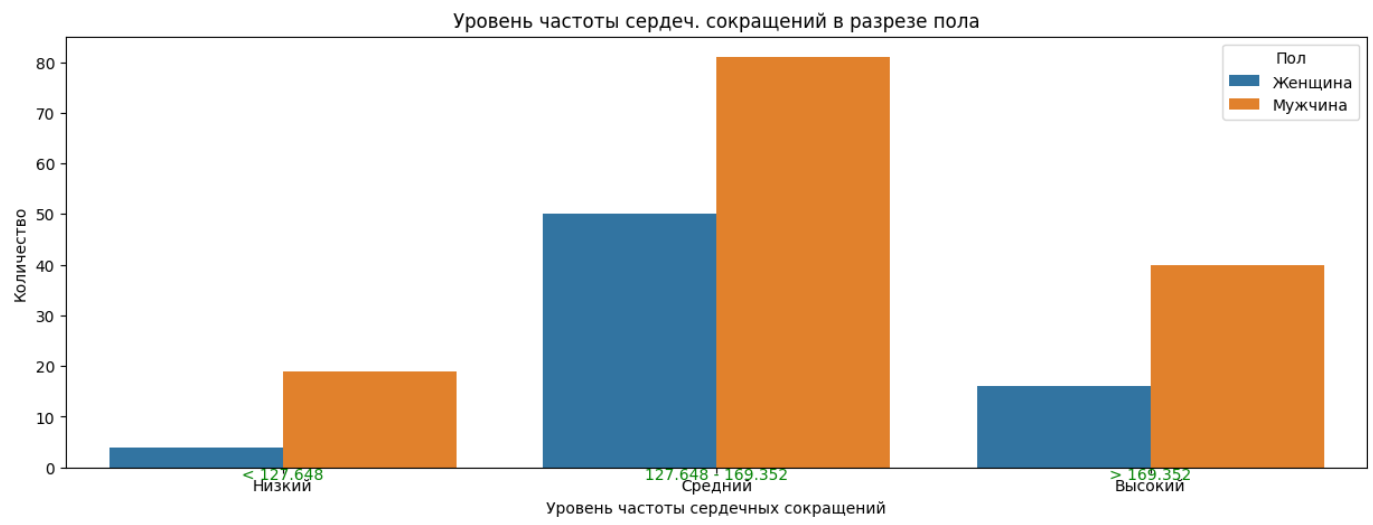
Большинство людей преобладают средний уровень частоты сердечных сокращений.

## Частота сердечных сокращений в разрезе пола

****

****

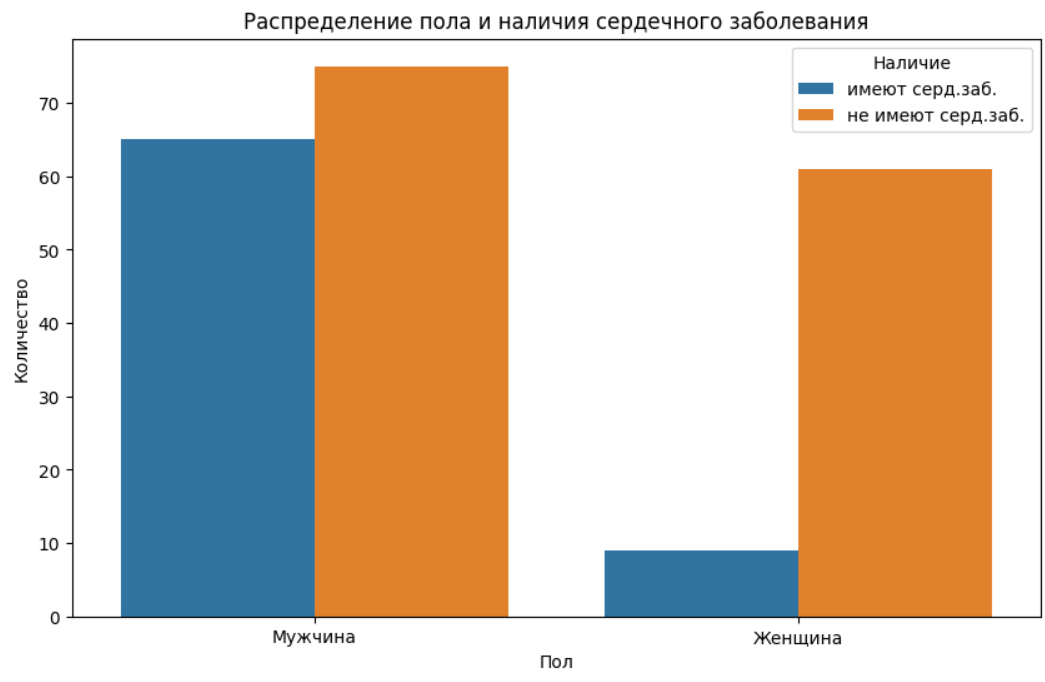
****

****

У мужчин частоты сердечных сокращений больше, чем у женщин.

## Распределение пола и наличия сердечных заболеваний

****

****

По графику видно, что большое количество у женщин и у мужчин не имеют сердечного заболевания, но у мужчин имеющиеся сердечные заболевания больше, чем у женщин.

Рассмотрим их на скрипичной диаграмме.

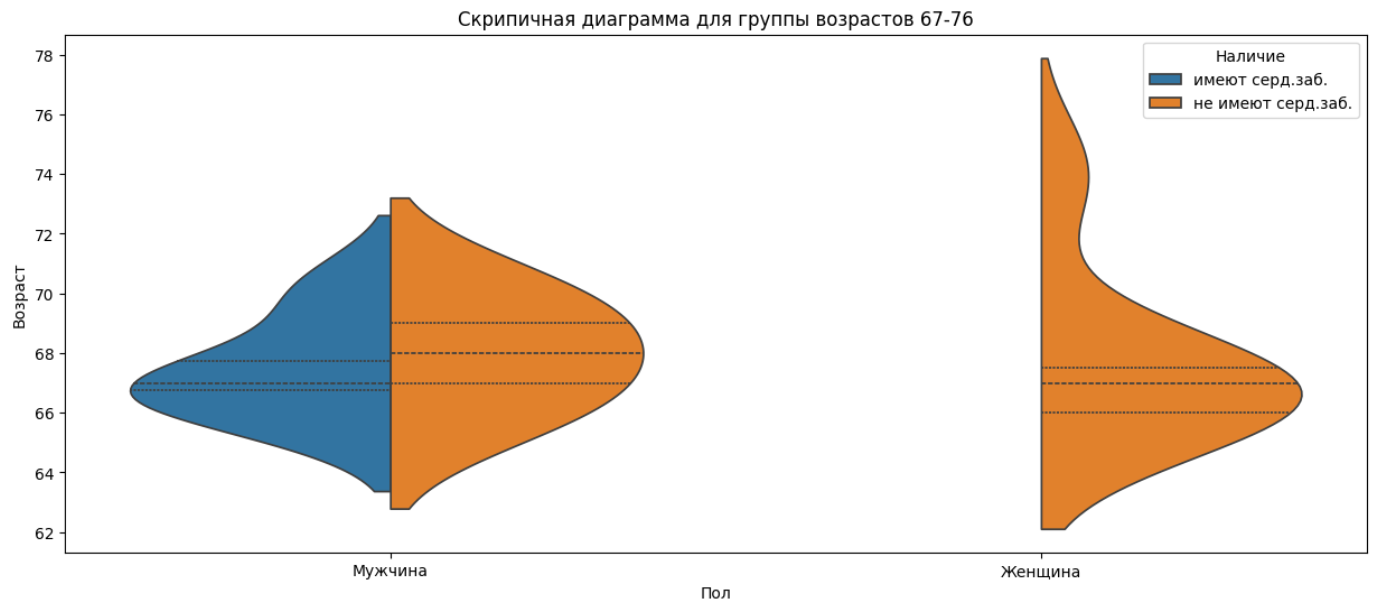
****

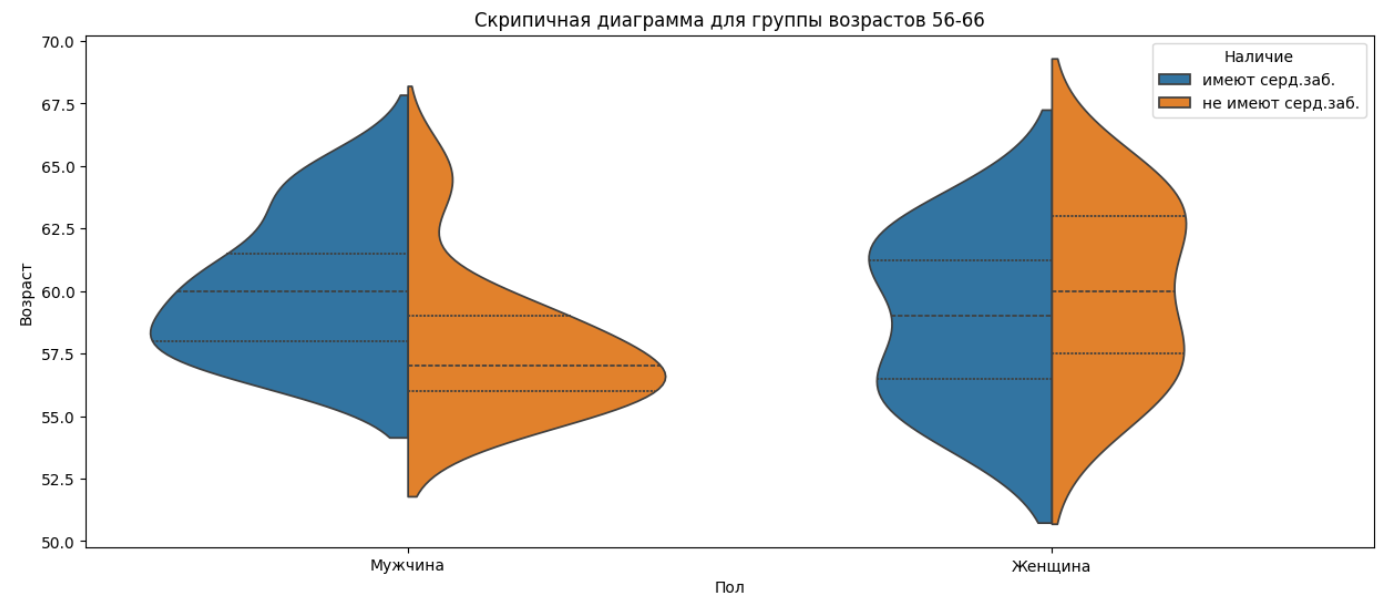
****

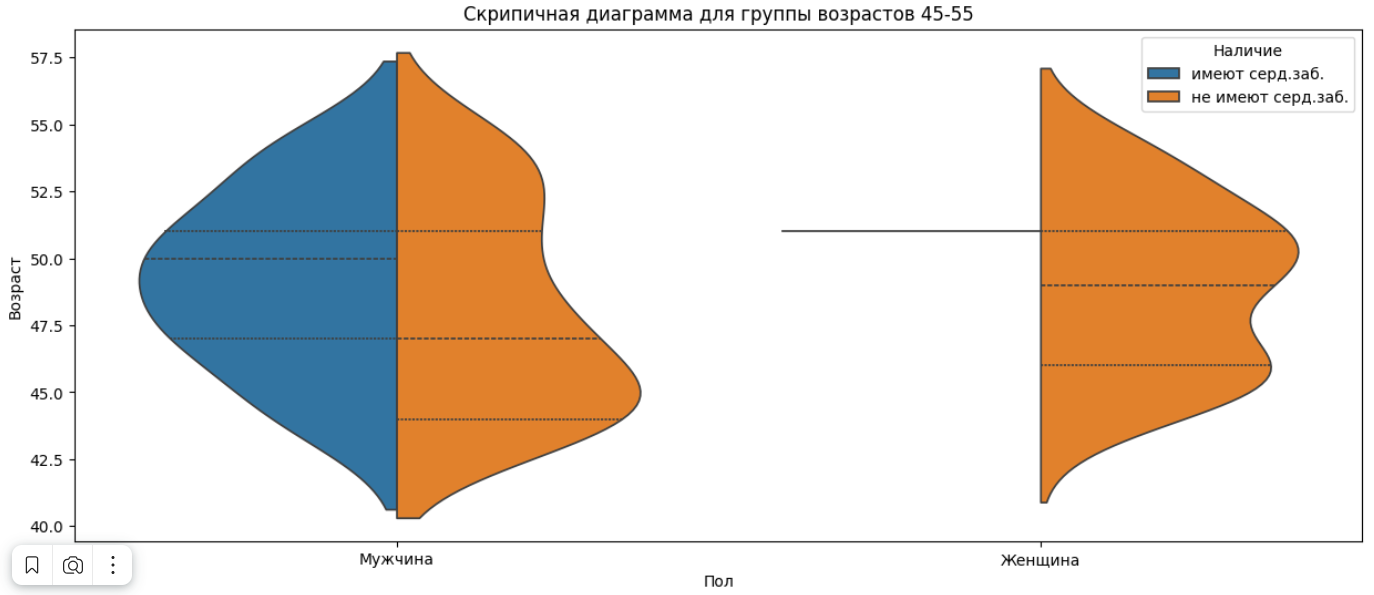
Женщины реже болеют, чем мужчины.

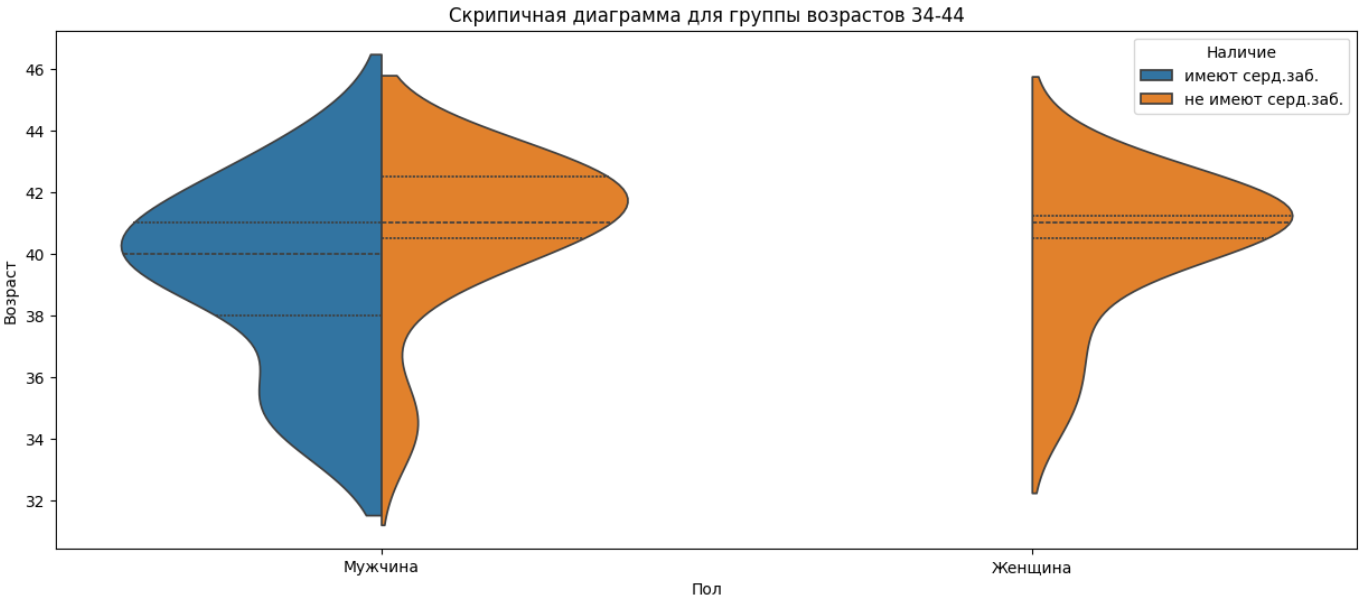
Рассмотрим по группам возрастов.











По графику видно, что меньше данных доступно для женщин с сердечными заболеваниями в возрастных группах 34-44, 45-55 и 67-76 лет. Поэтому рассмотрим количество значений наличия сердечных заболеваний у мужчин и женщин.



Исходя из подсчета количества значений target для каждого значения sex, можно сделать следующие выводы: Среди женщин меньше случаев сердечных заболеваний по сравнению с мужчинами. В то же время, среди женщин больше случаев, когда сердечные заболевания отсутствуют, чем среди мужчин. Эти выводы могут быть полезны при анализе рисков сердечных заболеваний в зависимости от пола.