ГУАП

КАФЕДРА № 41

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ассистент |  |  |  | В.В. Боженко |
| должность, уч. степень, звание |  | подпись, дата |  | инициалы, фамилия |

|  |
| --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ №2 |
| ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ |
| по курсу: ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ |
|  |
|  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СТУДЕНТ ГР. № | 4917 |  |  |  | В.С. Тихонов |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург

2022

**Цель работы:** осуществить предварительную обработку данных csv-файла, выявить и устранить проблемы этих данных.

Вариант 21. Ссылка на репозиторий: https://github.com/Deiwas/analysis\_lab2

1. Получен набор данных от преподавателя для проведения анализа согласно цели выполняемой работы (рис. 1).

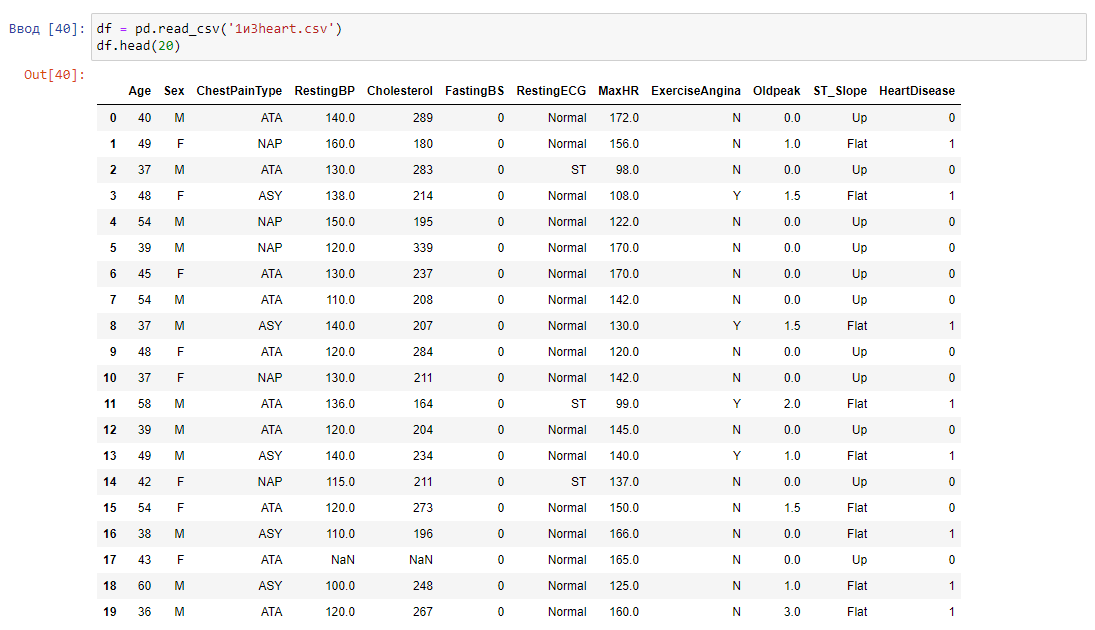


Рисунок 1 – Загруженный датафрейм

1. Проведена предварительная обработка данных. В процессе ее выполнения удалены дубликаты (рис. 2), пустые строки (рис. 3), проведена конвертация данных (рис. 4), а также проведена реиндексация датасета (рис. 5).

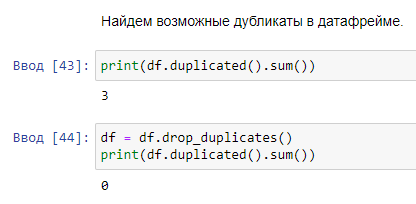


Рисунок 2 – Удаление дубликатов

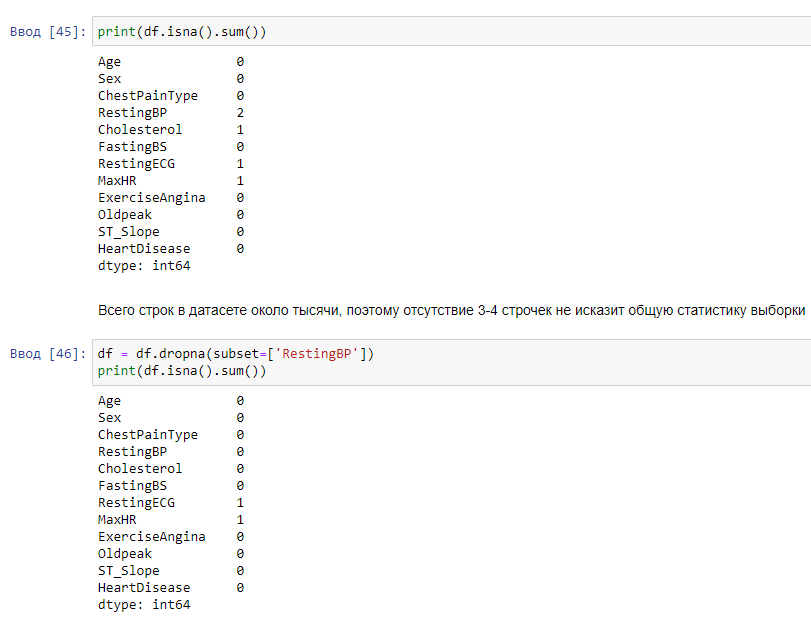


Рисунок 3 – Удаление пустых строк

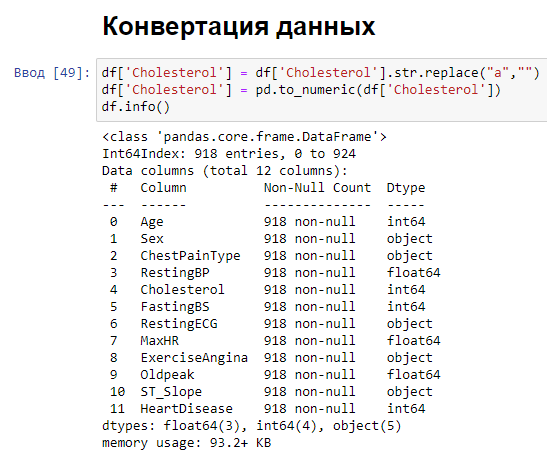


Рисунок 4 – Конвертация столбца «Cholesterol»

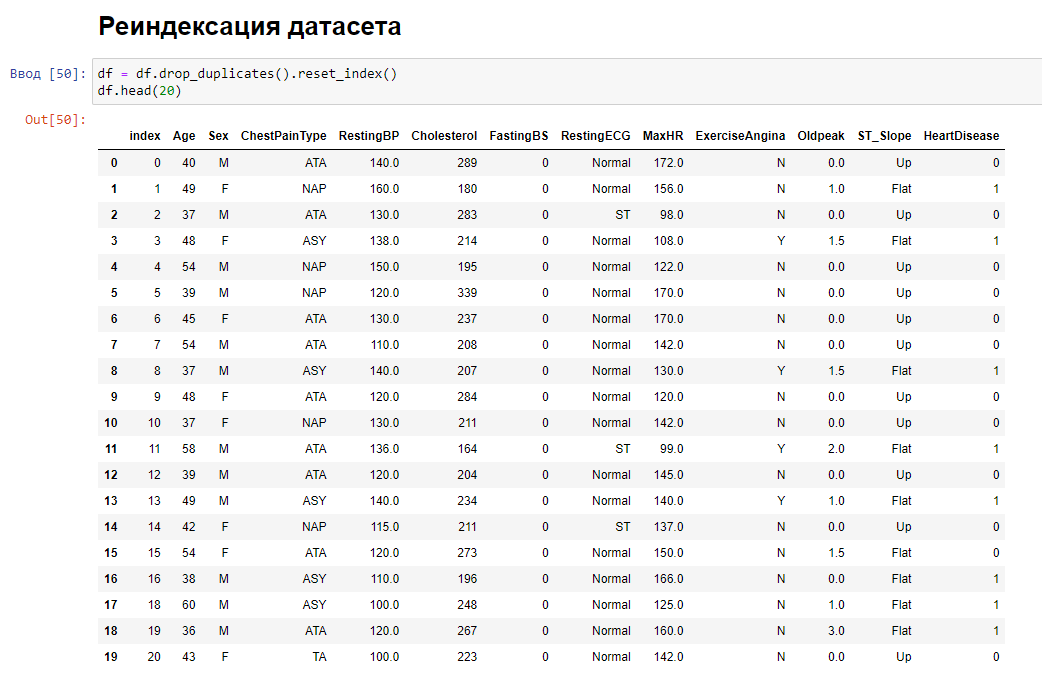


Рисунок 5 – Реиндексация датасета

1. Была построена точечная диаграмма для параметров «Age», «MaxHR» (рис. 6). Исходя из этой диаграммы, можно говорить об обратно пропорциональной зависимости этих параметров.

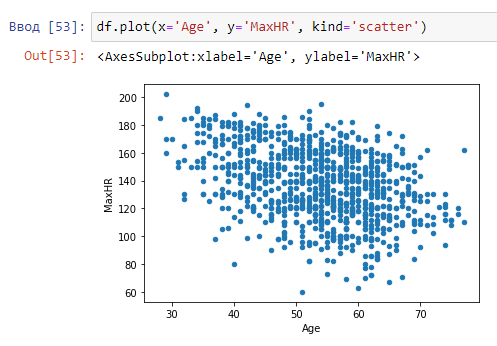


Рисунок 6 – Точечная диаграмма с обратной пропорциональностью

1. Была построена матрица диаграмм рассеивания для обнаружения зависимости между переменными датафрейма (рис. 7). Эта матрица подтверждает предположение о наличии только

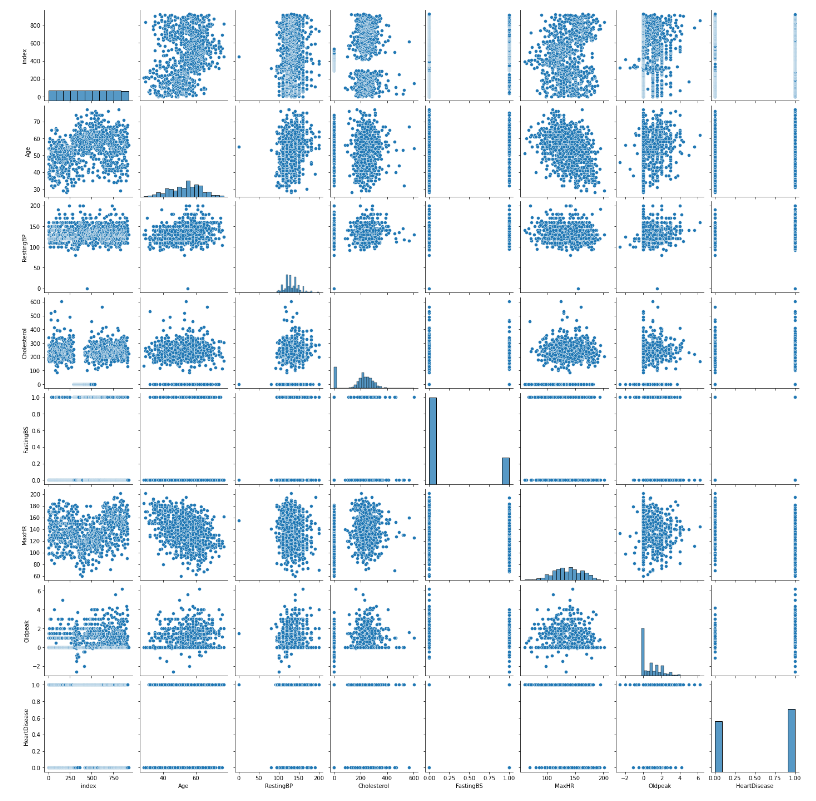


Рисунок 7 – Матрица диаграмм рассеивания

1. Вычислил корреляцию между различными столбцами с помощью команды corr() (рис. 8).

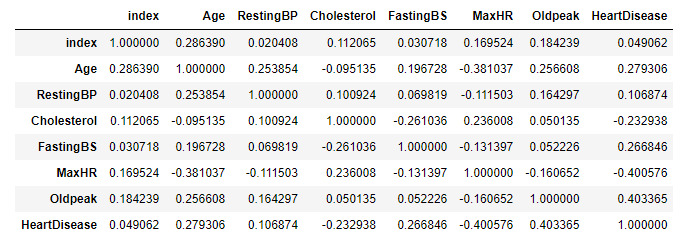


Рисунок 8 – Корреляция между различными данными

1. Была построена тепловая карта корреляции. В процессе постройки карты был опущен столбец «index» (рис. 8).

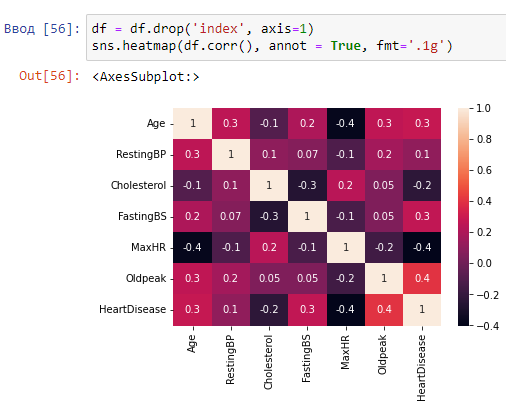


Рисунок 8 – Тепловая карта корреляции

1. Целевым признаком данной таблицы является столбец «HeartDisease». Это обусловлено тем, что все другие столбцы можно использовать для обучения модели тому, как выглядят параметры больных и здоровых людей.

**Вывод:**

В ходе выполнения данной лабораторной работы были получены навыки анализа данных на основе точечных диаграмм и корреляции данных. Для обработки при помощи библиотеки pandas был загружен набор данных о сердечных заболеваниях. После был проведен первичный анализ данных: удалены дубликаты, пустые строки и изменены неверные типы данных. В результате анализа данных с помощью точечных диаграмм была обнаружена обратно пропорциональная зависимость между возрастом и максимальным сердцебиением. Исходя из данных, это единственная зависимость в этом датасете. Это подтверждает и тепловая карта корреляции.

Список источников:

1. Методические указания к лабораторной работе №2: https://pro.guap.ru/get-task/12f07f5eb81ba1a11f3eb1a42897fbf0