**Московский государственный технический**

**университет им. Н.Э. Баумана**

Факультет «Информатика и системы управления»

Кафедра ИУ5 «Системы обработки информации и управления»

Курс «Технологии машинного обучения»

Отчет по рубежному контролю №1

«Технологии разведочного анализа и обработки данных»

Вариант №8

|  |  |  |
| --- | --- | --- |
| Выполнил: |  | Проверил: |
| студент группы ИУ5-61Б |  | преподаватель каф. ИУ5 |
| Коваленко Алексей Викторович |  | Гапанюк Юрий Евгеньевич |
|  |  |  |
| Подпись:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Дата:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  | Подпись:\_\_\_\_\_\_\_\_\_\_\_\_\_  Дата:\_\_\_\_\_\_\_\_\_\_\_\_\_ |
|  |  |  |
|  |  |  |

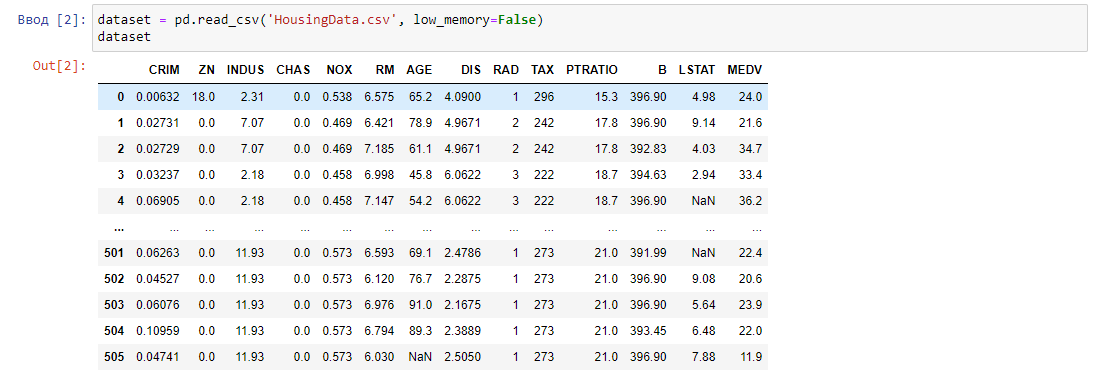
Москва, 2023 г.

**Выполнение работы**

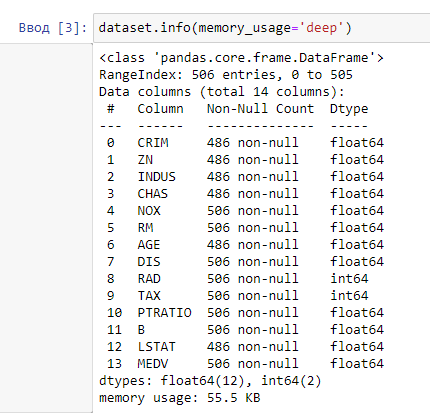
Для выполнения задачи проведения корреляционного анализа был взят набор данных «Boston housing dataset», который был предварительно загружен с ресурса kaggle.com.



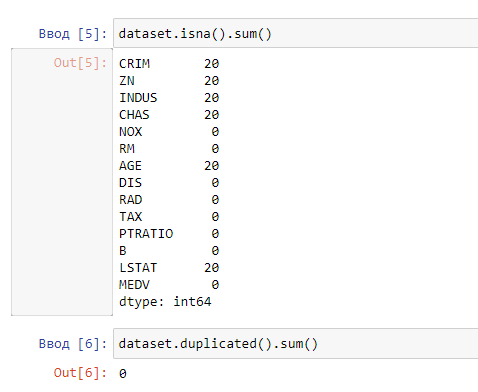
Для проведения анализа был создан датафрейм, который содержит 13 нецелевых признаков (параметры жилья в Бостоне) и 1 целевой (стоимость жилья в Бостоне)



Все поля имеют числовой тип, среди которых есть как целочисленные данные, так и дробные.



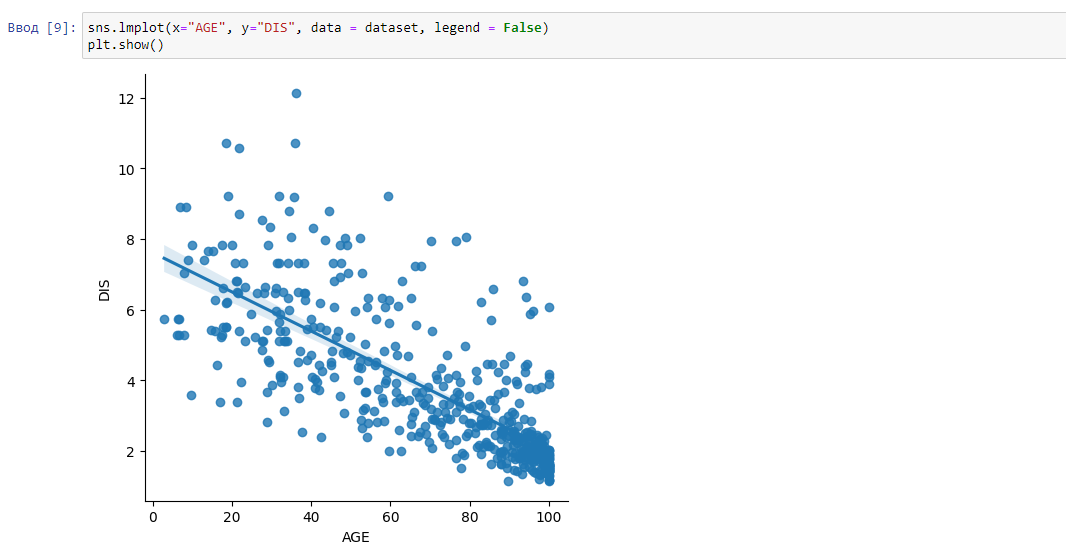
В наборе данных присутствуют пропуски и отсутствуют дубликаты.



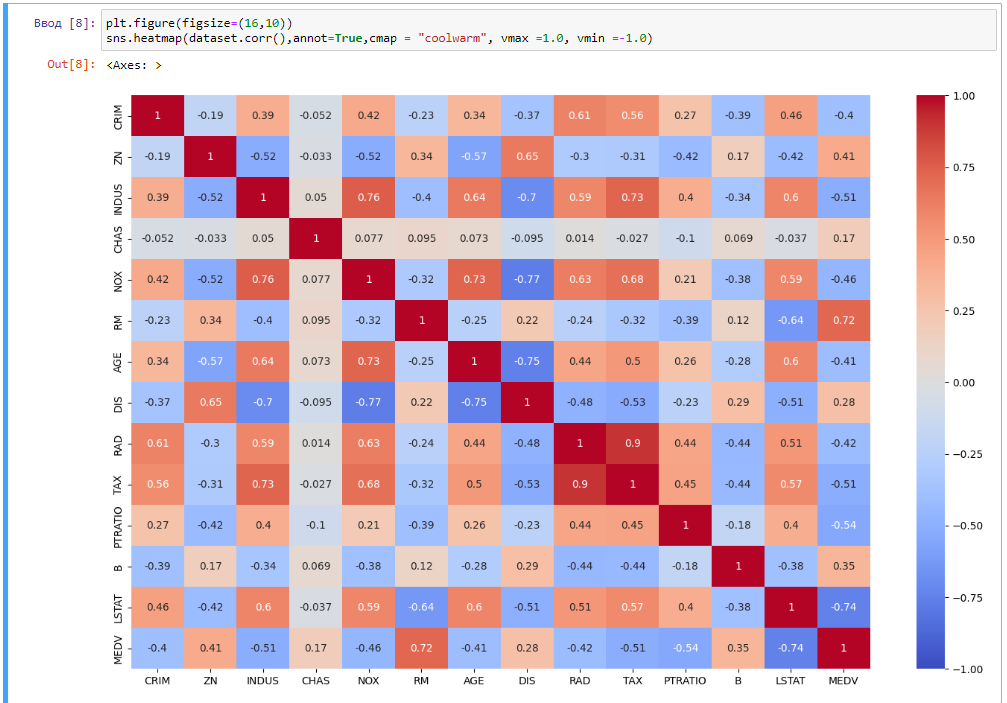
Для обработки пропусков удалим строки с пропусками.



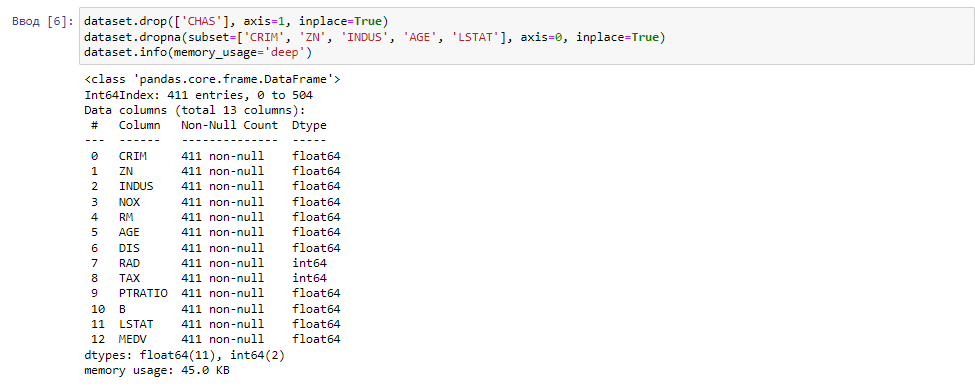
Для пары колонок AGE и DIS был построен график “Диаграмма рассеяния”.

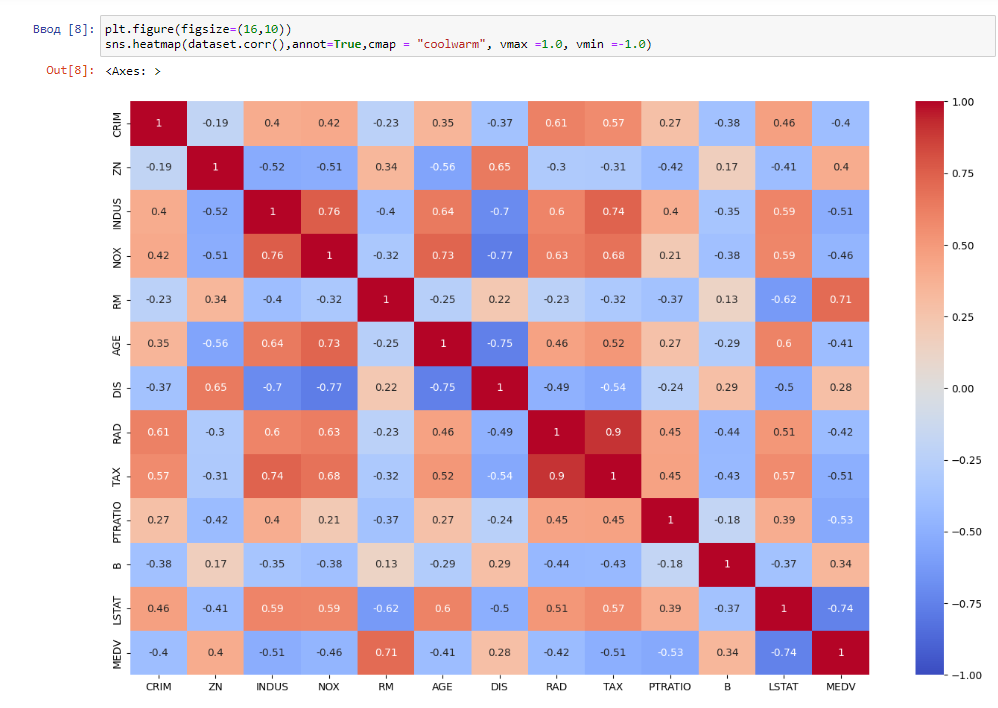


Для визуализации корреляционной матрицы была использована “тепловая карта”.



На основе данной диаграммы мы можем сразу сказать, что есть признак, который не коррелирует с другими признаками полностью, а также на шаге очистки данных, мы удаляли строки с пропусками данных в данном столбце. Так что, основываясь на полученном результате, можно принять решение, что можно удалить данный столбец еще на этапе очистки данных, что позволит сохранить часть строк, где содержались полезные данные.





По полученной диаграмме можно сделать следующие выводы.

С целевым признаком наибольшую корреляцию имеют признаки:

С целевым признаком наиболее сильную корреляцию имеют признаки “RM” (0,71) и “LSTAT” (-0,74). Эти признаки являются наиболее информативными при работе с построением моделей. Некоторые другие признаки имеют более низкую корреляцию, но которая все еще оказывает воздействие: “PTRATIO” (-0,53), “TAX” (-0,51), “INDUS” (-0,51), “NOX” (-0,46), “AGE” (-0,41), и “RAD” (-0,42). Эти признаки все еще стоит использовать при создании модели. Оставшиеся признаки, такие как “CRIM” (-0,4), “ZN” (0,4), “DIS” (0,28) и “B” (0,34), не стоит использовать из-за малого влияния на целевой признак.

Таким образом мы получаем, что на основе признаков “RM”, “LSTAT”, PTRATIO”, “TAX”, “INDUS”, “NOX”, “AGE”, и “RAD” можно построить модель для машинного обучения, где первые 2 признака внесут наибольший вклад в полученную модель.