**Отчёт по домашнему заданию 2.2**

**“Задание с Colab”**

Выполнил:

студент БПМ-17-2

Скибин К.С.

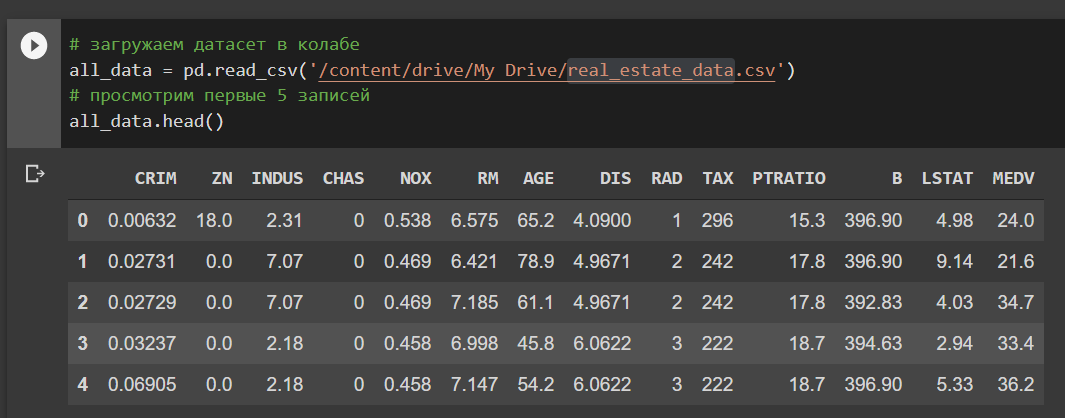
**Описание задачи**

Дан датасет real\_estate\_data, содержащий информацию об уровне преступности на душу насления в пригородах Бостона, а также общую информацию о стоимости жилья с учетом различных социальных факторов.

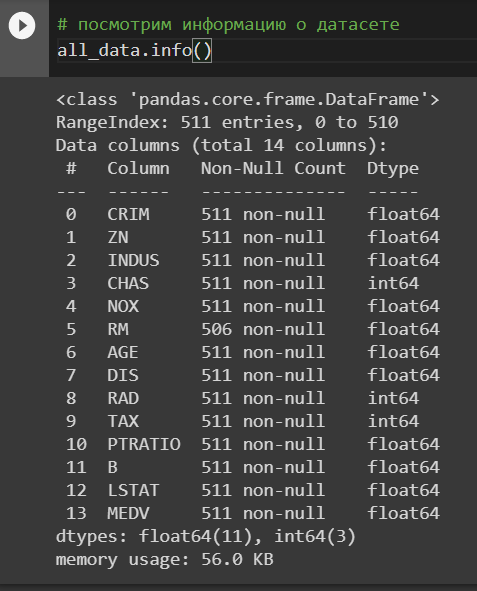
Необходимо создать модель, лучше всего предсказывающую уровень преступности по остальным параметрам

**Загрузка, первичная обработка датасета**

Загружаем датасет. Выводим первые 5 записей.



Посмотрим подробную информацию о полях и количестве записей.

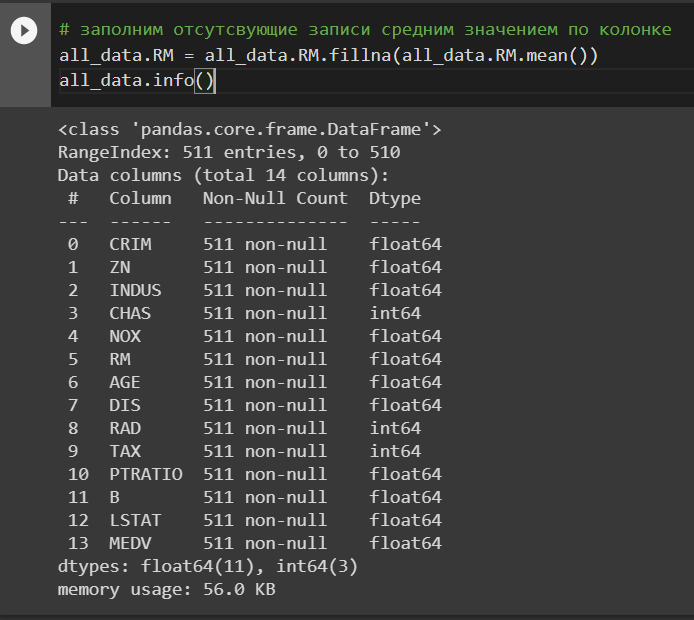


Всего записей в таблице – 511.

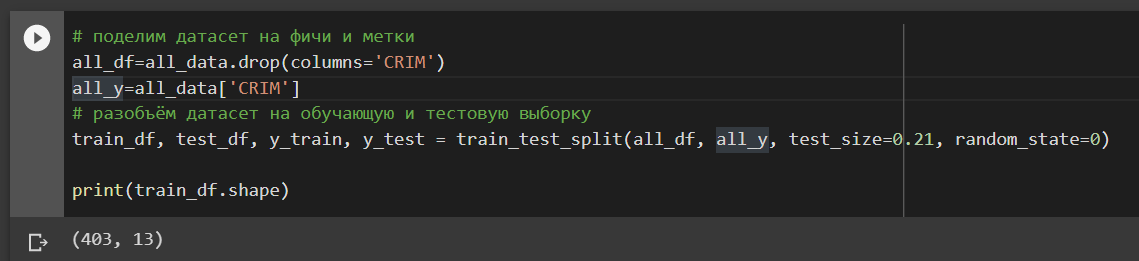
**Описание колонок:**

* CRIM - уровень преступности на душу населения
* ZN - доля коммерческой собственности
* INDUS - доля не розничного бизнеса
* CHAS - фиктивная переменная Charles River (= 1, если участок граничит с рекой; 0 в противном случае)
* NOX - концентрация оксидов азота
* RM - среднее количество комнат в доме
* AGE - доля жилья, построенного до 1940 г.
* DIS - взвешенные расстояния до пяти Бостонских центров занятости
* RAD - индекс доступности радиальных магистралей
* TAX - полная ставка налога на имущество за 10 000 долларов США
* PTRATIO - количество учеников на одного учителя
* B - 1000(Bk - 0.63)^2 где Bk доля темнокожего населения
* LSTAT - доля населения с низким социальным статусом
* MEDV - средняя стоимость домов, занимаемых владельцами, в тысячах долларов

Колонка RM содержит 5 пропущенных записей. Заполним их средним значениями по всей колонке RM.

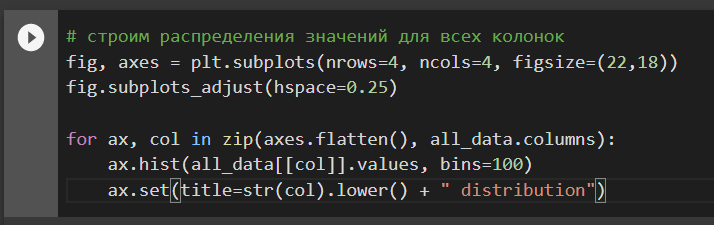


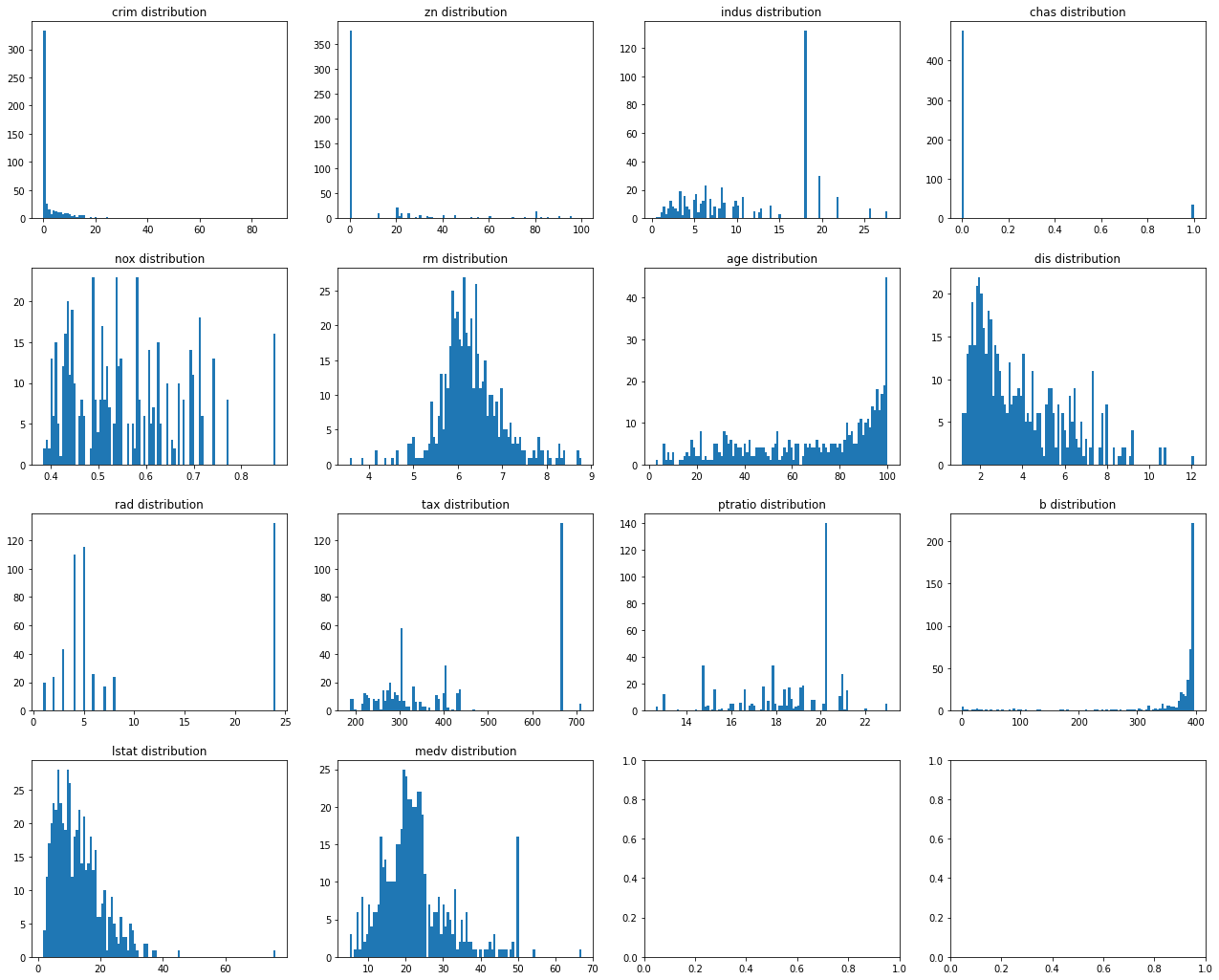
Разобъём датасет на фичи и метки, убирая и выделяя колонку CRIM соответственно. Далее разобъем полученные массивы на тренировочную и тестовую подвыборки.



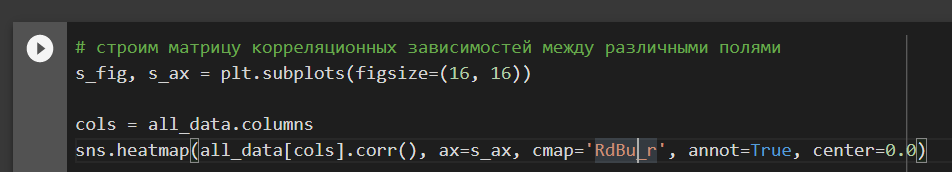
**Разведочный анализ данных**

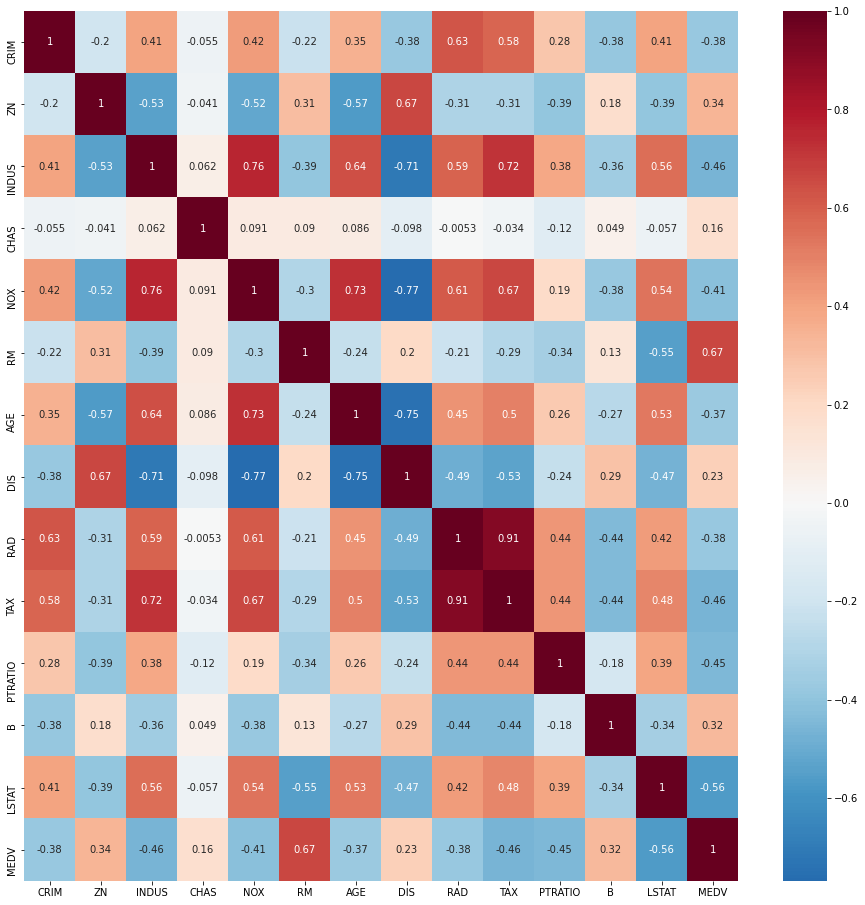
Построим частотные распределения значений для всех колонок





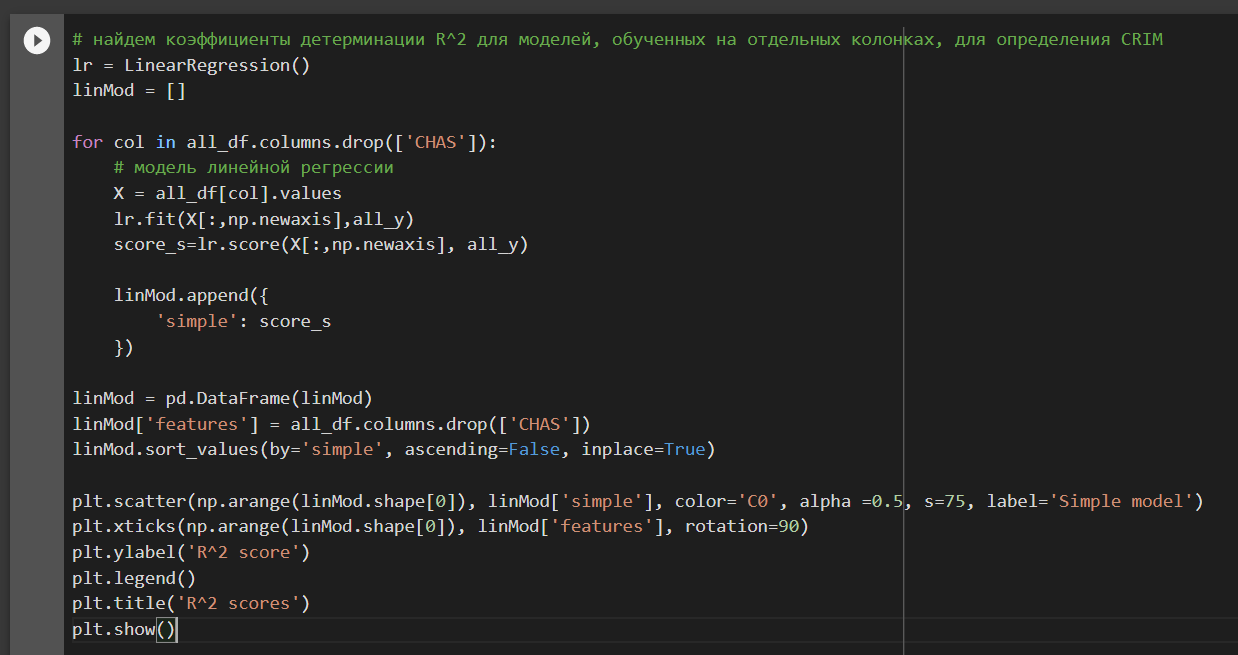
Построим матрицу коэффициентов корреляции между различными полями

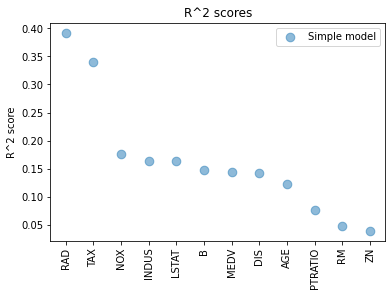




В первой строке матрице можно увидеть корреляцию между уровнем преступности и всеми остальными параметрами. Можно увидеть, что уровень преступности практически не зависит, граничит ли собственность с рекой (CHAS)

Найдем коэффициенты детерминации R^2 для моделей, обученных на отдельных колонках, для определения CRIM



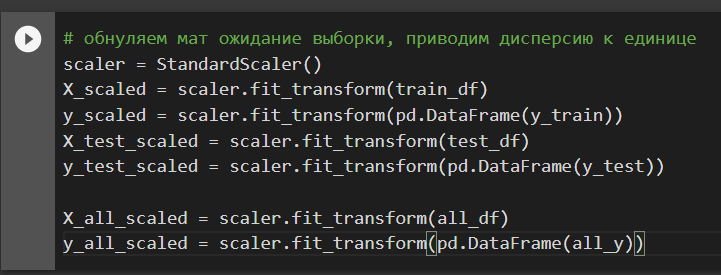


Как мы видим наибольшая зависимость прослеживается между CRIM и полями RAD (индекс доступности радиальных магистралей) и TAX (полная ставка налога на имущество за 10 000 долларов США).

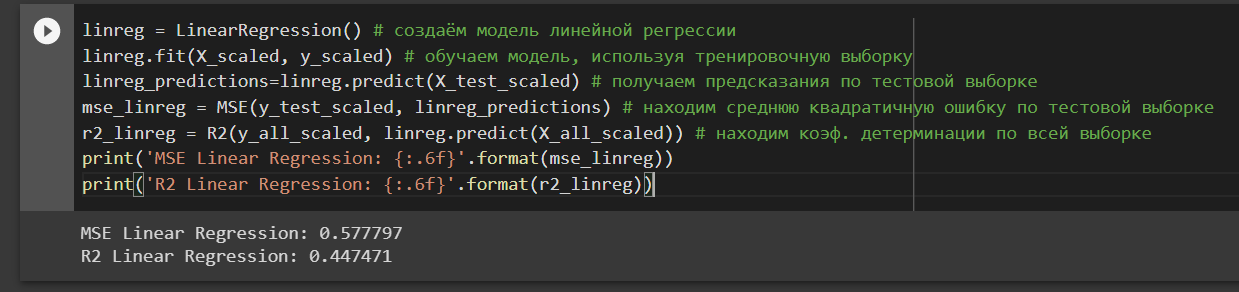
Наименьшее влияние на CRIM оказывают ZN (доля коммерческой собственности), PTRARIO (количество учеников на одного учителя) и RM (среднее количество комнат в доме)

**Тренировка и оценка линейных моделей**

Для начала отмасштабируем все выборки, чтобы мат. ожидание равнялось 0, а дисперсия 1.



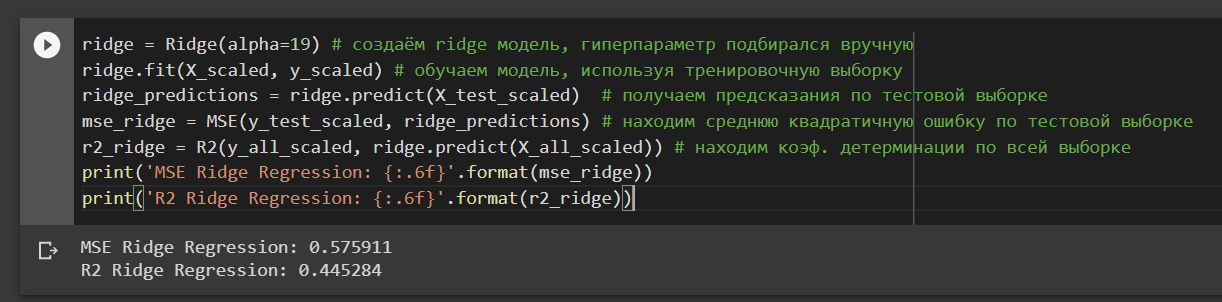
Построим модель линейной регрессии. Обучим её на тренировочной выборке. Оценим среднюю квадратичную ошибку на тестовой выборке. Найдём коэффициент детерминации по всей выборке.



Средняя квадратичная ошибка: ~0.577797

Коэффициент детерминации: ~0.446471

Построим модель Ridge модель. Обучим её на тренировочной выборке. Оценим среднюю квадратичную ошибку на тестовой выборке. Найдём коэффициент детерминации по всей выборке. Подберём целый гиперпараметр, чтобы модель давала наименьшую ошибку.



Средняя квадратичная ошибка: ~0.575911

Коэффициент детерминации: ~0.445284

**Заключение и результаты**

В результате проведенного разведочного анализа данных были сделаны выводы:

* Уровень преступности (CRIM) практически не зависит, граничит ли собственность с рекой (CHAS)
* Наибольшая зависимость прослеживается между CRIM и полями RAD (индекс доступности радиальных магистралей) и TAX (полная ставка налога на имущество за 10 000 долларов США)
* Наименьшее влияние на CRIM оказывают ZN (доля коммерческой собственности), PTRARIO (количество учеников на одного учителя) и RM (среднее количество комнат в доме)

В целом прослеживались прямые и обратные линейные зависимости между параметрами и уровнем преступности. Максимальный коэффициент детерминации наблюдался у модели линейной регрессии с параметром RAD (индекс доступности радиальных магистралей) - 0.39. Было принято решение построить две модели, задействующие все параметры:

* Модель линейной регрессии
* Ridge модель (регрессия с использованием метода регуляризации Тихонова)

Оценить их коэффициенты детерминации на всей выборке и средние квадратичные ошибки на тестовой выборке и сравнить между собой.

При построении Ridge модели осуществлялся ручной подбор гиперпараметра альфа с целью получить минимум по MSE. В итоге альфа = 19.

Полученные итоговые результаты:

1. MSE Linear Regression: 0.577797
2. MSE Ridge Regression: 0.575911
3. R2 Linear Regression: 0.447471
4. R2 Ridge Regression: 0.445284

Ridge модель незначительно превосходит модель линейной регрессии по ошибке ну уступает по коэффициенту детерминации. При этом обе модели смогли превзойти линейную регрессию с одним параметром RAD по коэффициенту детерминации приблизительно на 13%