**Московский государственный технический   
университет им. Н. Э. Баумана**

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №1

«Технологии разведочного анализа и обработки данных.»

Вариант № 16

Выполнил: Проверил:  
Пронченко Т.А. Гапанюк Ю.Е.  
группа ИУ5-62Б

Дата: 14.04.25 Дата:

Подпись: Подпись:

Москва, 2025 г.

**Полученное задание**

Номер варианта: **16**

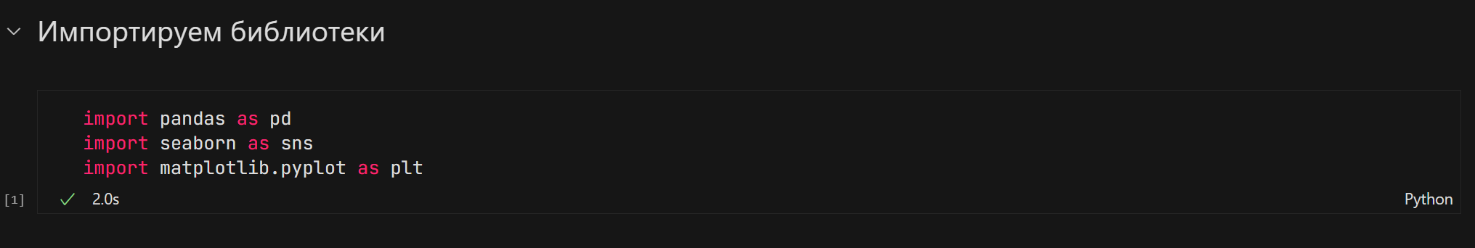
Номер задачи: **2**

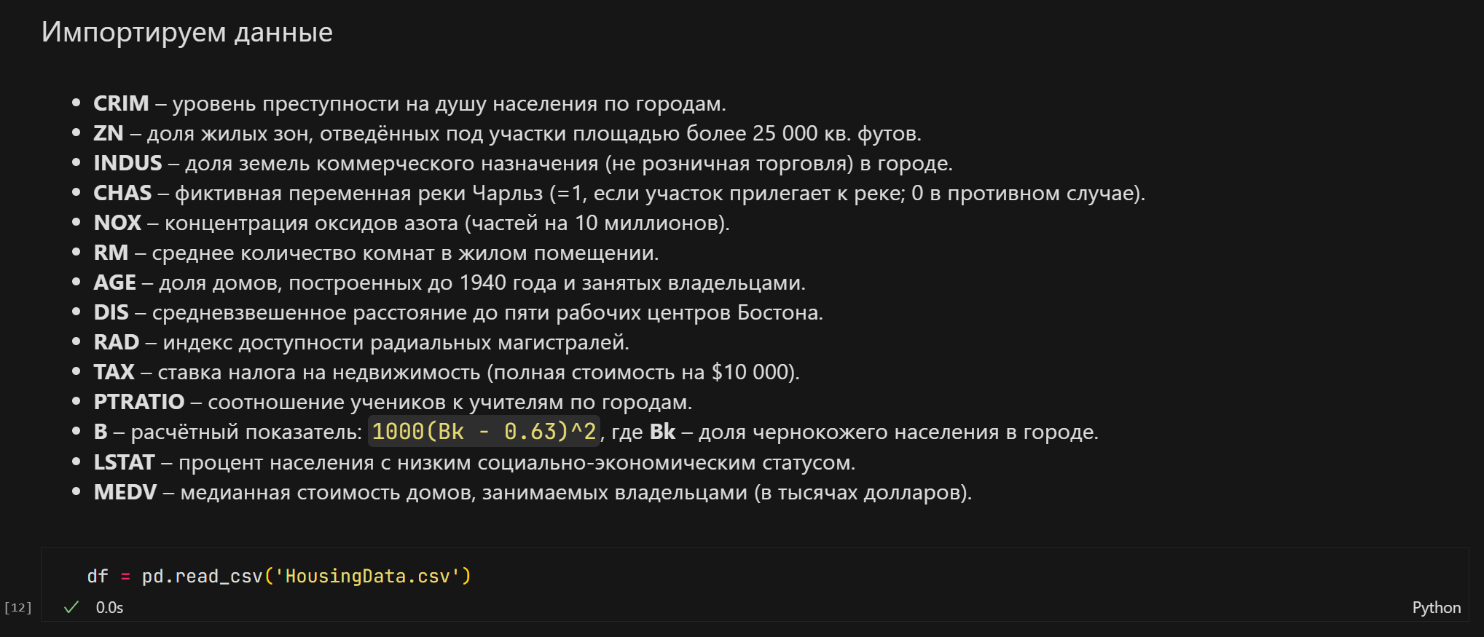
1. Номер набора данных, указанного в задаче: **8** <https://www.kaggle.com/datasets/altavish/boston-housing-dataset>

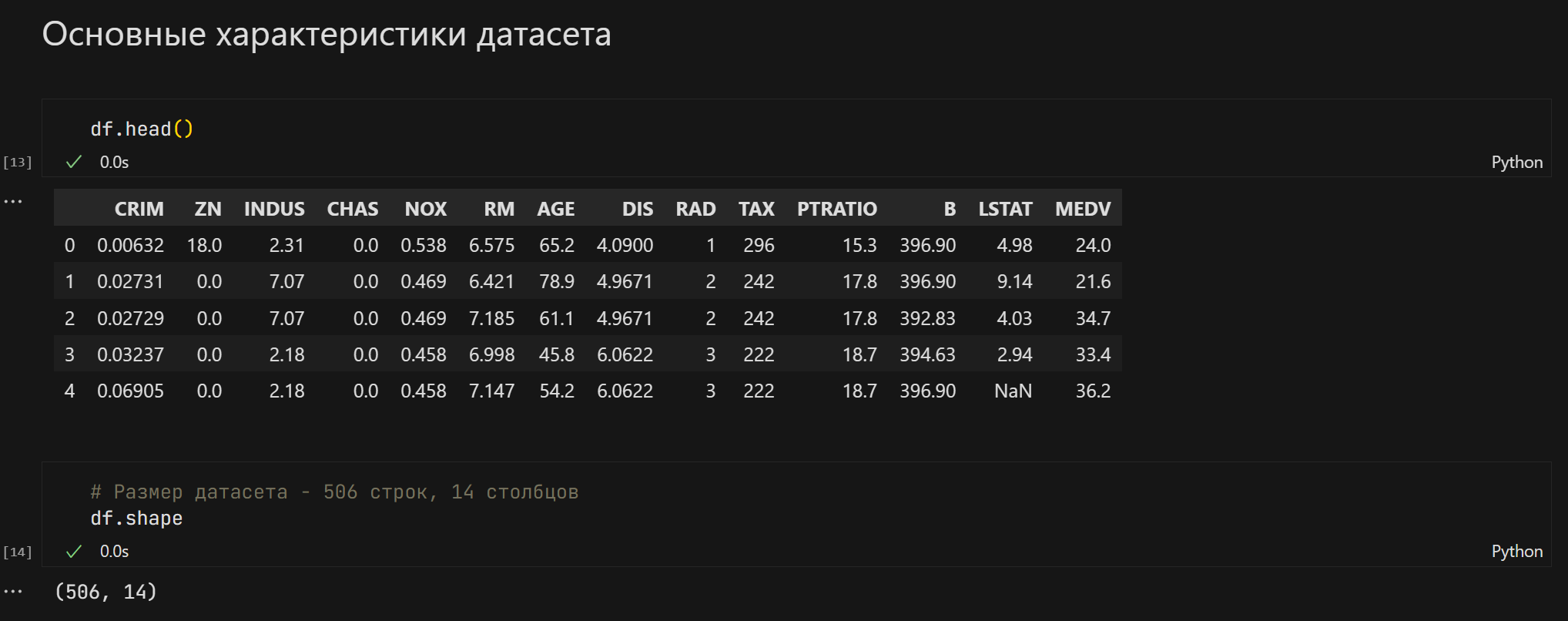
**Задача №2.**

Для заданного набора данных проведите обработку пропусков в данных для одного категориального и одного количественного признака. Какие способы обработки пропусков в данных для категориальных и количественных признаков Вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

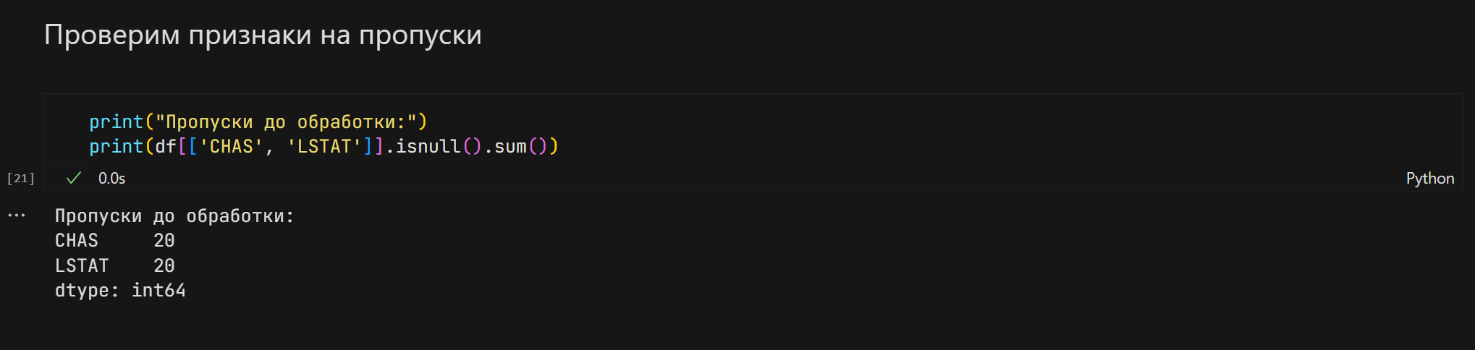
**Ход выполнения**

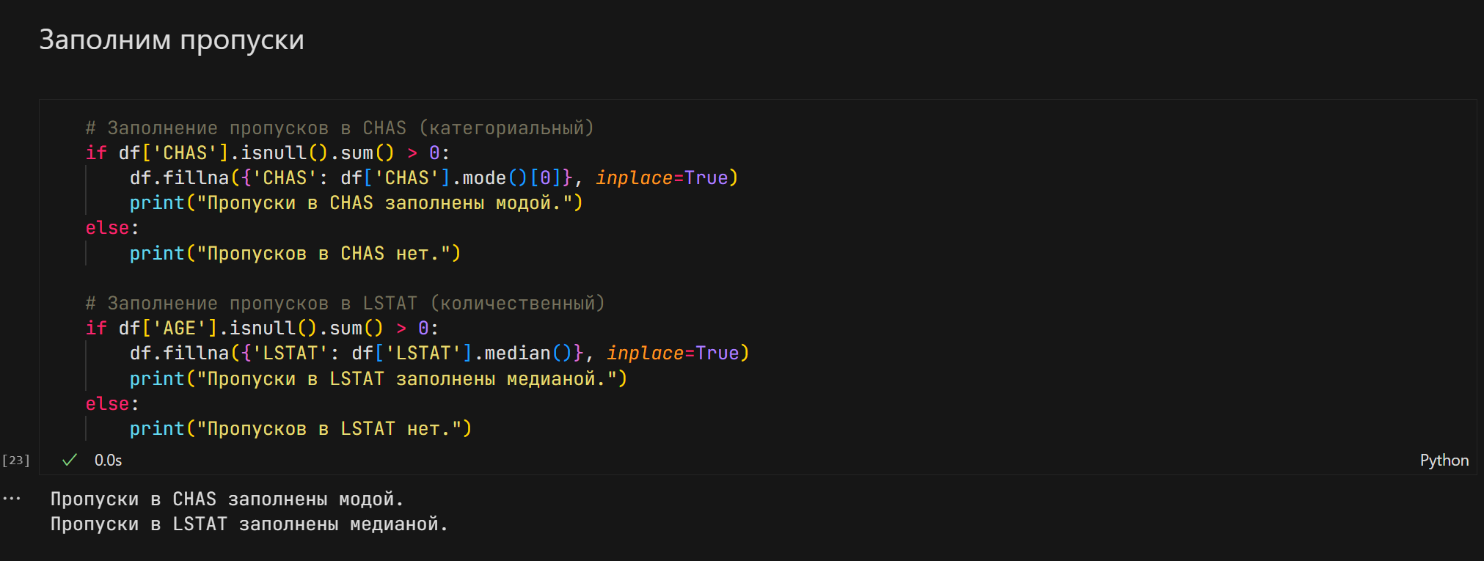
****

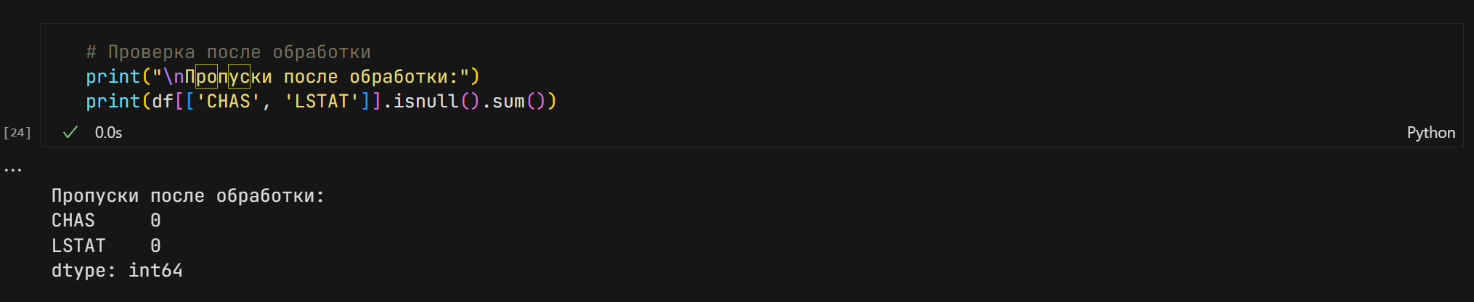




В качестве категориального признака выберем бинарный признак «CHAS» - переменная реки Чарльз, указывающая прилегает ли участок к реке. В качестве количественного признака выберем признак «LSTAT» - процент населения с низким социально-экономическим статусом.







**Какие способы обработки пропусков в данных для категориальных и количественных признаков вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?**

1) Признак CHAS является бинарной переменной, принимающей значения 0 или 1, где 1 обозначает расположение недвижимости вблизи реки Чарльз, а 0 - отсутствие такого соседства. Для обработки пропущенных значений в данном признаке могут быть применены следующие методы:

1. **Замена на моду (наиболее частое значение)**  
   Данный подход целесообразен при небольшом количестве пропусков и значительном преобладании одного из значений.
2. **Создание новой категории "Unknown"**  
   При существенном количестве пропусков рекомендуется введение дополнительной категории.
3. **Предсказание пропущенных значений**  
   При наличии значительного числа пропусков возможно применение простых моделей для прогнозирования отсутствующих значений на основе других признаков.

Т.к. в нашем случае пропуски составили менее 5% и значение 0 преобладает замена на моду будет самым оптимальным способом.

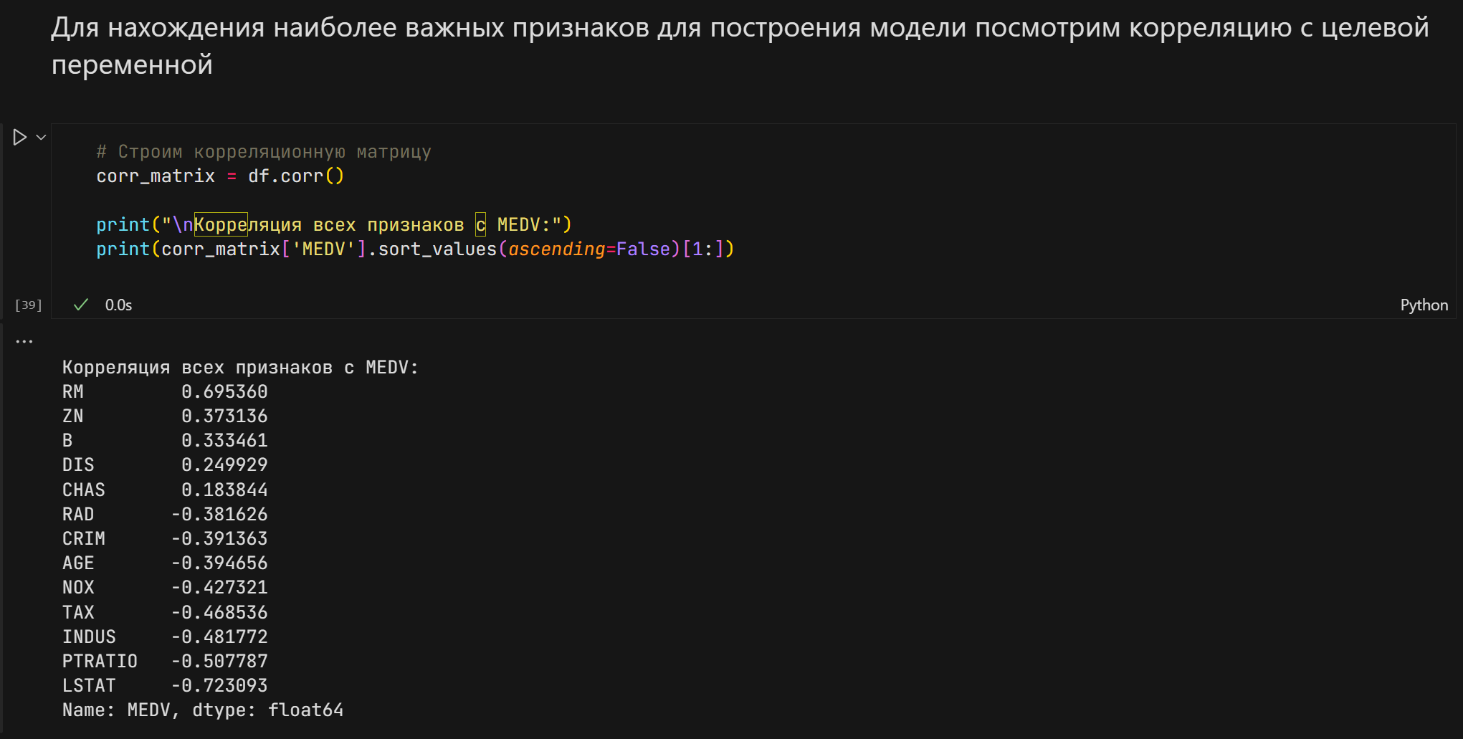
2) Признак LSTAT отражает процент населения с низким социально-экономическим статусом. Методы обработки пропусков:

1. **Замена на медиану**  
   Медиана является устойчивой к выбросам мерой центральной тенденции
2. **Замена на среднее значение**  
   При нормальном распределении данных и отсутствии выбросов допустимо использование среднего
3. **Замена на константу**  
   В некоторых случаях применяется заполнение фиксированным значением (0 или -1), однако это может исказить данные.
4. **Использование KNN Imputer**  
   Для большого числа пропусков эффективен алгоритм k-ближайших соседей:

Т.к. в нашем случае пропусков немного, и можно заметить выбросы на диаграмме, наиболее оптимальным способом будет замена на медиану.

3) Выбор признаков для построения модели машинного обучения

Для датасета Boston Housing (CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT, MEDV) отбор наиболее информативных признаков осуществлялся на основе корреляции с целевой переменной (MEDV - стоимость жилья).



**Наиболее значимые признаки:**

* **LSTAT** (% населения с низким статусом): обратная зависимость от стоимости
* **RM** (среднее число комнат): сильная положительная корреляция с ценой
* **PTRATIO** (соотношение учеников и учителей): влияет на привлекательность района
* **INDUS** (доля нежилых площадей): обратная зависимость от стоимости - чем больше промышленных зон, тем ниже цены на жилье
* **TAX** (налог на имущество): высокая налоговая нагрузка снижает стоимость недвижимости

Несмотря на то, что некоторые признаки выделяются на фоне других, в нашем случае для построения моделей машинного лучше использовать все признаки, т.к. каждый из них все равно оказывает значительное влияние на стоимость участка.